The Effect of Reward Systems on Gameplay

Frank Bruni
UC Berkeley MIDS
frankbruni@berkeley.edu

Casey Yoon
UC Berkeley MIDS
caseyhyoon@berkeley.edu

Ash Tan
UC Berkeley MIDS
asht@berkeley.edu

A. Abstract

Many mobile apps and video games implement a system that rewards consumers for using the app/playing the game, often in the guise of something like a level or point system. For example, the longer a consumer uses an app or plays a game, they might be rewarded with more points or a level-up in order to motivate them to keep using the app/playing the game. These incentives are often intended to provide the player with positive experiences that promote increased engagement, and this style of reward system has even been implemented in educational applications in efforts to increase student engagement. The prevalence of this phenomenon raises some interesting questions. Do these reward systems cause consumers to use these apps more often? If so, does knowledge of these intended effects affect the efficacy of a reward system? In our study, we found no statistically significant evidence that reward systems affect gameplay time or engagement, and no statistically significant evidence that knowledge of the intended reward system affects gameplay time or engagement.

B. Background

Reward systems are becoming increasingly prevalent in modern services. Understanding how consumers react to these incentive systems is crucial to understanding how these systems could shape app development and online services going forward. Can the success of certain apps/games be attributed to the reward system they implement, or are they successful because people like the actual content? If it turns out that reward systems are effective at capturing attention even if the intent is known, then we might want to ask ourselves whether this strategy is simply an effective tool or an exploitative one. Furthermore, if knowledge of reward systems affects player behavior, information campaigns could be an effective tool to combat potentially exploitative reward systems.

C. Research Question

Do reward systems increase consumer engagement? If so, does knowledge of these intended effects affect the efficacy of a reward system?

D. Hypothesis

a. We hypothesize that implementing reward systems in games causes users to play and interact with the game more. We believe that reward systems will incentivize players to play more and with greater interest. These interactions will be tracked by time played and number of clicks.

b. During our randomization inference we also use a sharp null hypothesis. The sharp null hypothesis is such that the implementation of a reward system has no effects whatsoever on the user.

E. Research Design

E.i Experiment Overview

The experimental design itself consisted of a simple, intuitive, online puzzle game based on the "Linjat" game by Snellman (https://linjat.snellman.net/). We created three versions of this app: the control group, the first treatment group, and the second treatment group. The control group is given a version of the game with no implemented reward system. Users in the control group were only able to complete puzzles with no feedback, points, or rewards given in return. Both treatment groups received a version of the game with an implemented reward system: for each puzzle the user completes, they receive some number of experience points, meaning that the longer a user plays the more points they accrue (The puzzles did not increase in difficulty, as perceived difficulty could differ from individual to individual and would add additional and unnecessary complexity to the experiment, since the puzzles themselves have no bearing on the implemented reward system.) As an added reward, reaching certain point thresholds enables the user to change fonts on the game, with each point threshold unlocking an additional font option. This system is intended to motivate playing by creating a sense of progression. The difference between the two treatment groups is that the first treatment group receives a disclaimer notification before being able to play, while the second treatment group does not receive any disclaimer. This disclaimer notifies the subjects in the first treatment group that the reward system is intended to motivate them to play the game. This disclaimer allows us to study whether or not knowing the intended effect of the reward system affects consumers' behavior.

Upon loading into the game's landing page, the page itself randomly chooses the player's group (control/treatment 1/treatment 2) and, once the player agrees to the terms and conditions of the experiment, redirects them to the appropriate version of the game. User behavior is collected unobtrusively using Google Analytics and analyzed using R/RStudio.

E.ii Project Timeline

	Task	**Timeline**
Phase 1	Create website/game	2 weeks
Phase 2	Set up google analytics	1 week
Phase 3	Deploy website and find subjects	2 weeks
Phase 4	Analyze the experiment	2 weeks

E.iii Variable Measurement

The app, in all three groups, tracks the subject's activity on the app using Google Analytics. We set up a Google Analytics page for each group separately to ensure our data collection stayed pure. For each site we set up a system to track the duration of user time spent on the site, the number of total clicks by the user, number of clicks on the "Done" button by the user, and the number of clicks on the "Font" button by the user. Our reasoning behind this was to have two outcomes to test treatment effect, both duration and clicks. We define "engagement" to be the combination of the user's time spent and number of clicks spent on the site. The combined outcome variable handles outliers where users frantically click buttons and spend little time on the site and those users who have the site open but are inactive. Prior to combining, we log the two variables because the data is heavily skewed right to meet our normality assumption before implementing a linear regression model.

Google Analytics provided this data on each user in each group. A total of 119 participants were broken down into 61 in control, 31 in treatment 1 (with disclaimer), 27 in treatment 2 (without disclaimer).

E.iv Recruitment Process

Given limitations to data collection in Google Analytics, we were less inclined to utilize demographic information and instead took a naive approach to our recruitment process and found that anonymous subjects from popular online platforms were best for our field experiment. The recruitment process was performed online by posting the link to the study webpage (https://ashqtan.github.io/testing.github.io/) on various online platforms such as Reddit, Discord, and other social media sites. Discord servers include Gaming at Berkeley and Spring Bears (/r/Berkeley discord server). Participants were offered the chance to win $25 Amazon giftcards as a reward for participating in the experiment. Given our framework, this study had no attrition, and results from all participants were used.

Advertising Channels	
SubReddit channels	r/experiments, r/SampleSize, r/Berkeley, r/Puzzlevideogames, r/playmygame, r/surveys
Discord servers	Gaming at Berkeley

Given the lack of preceding research, we were forced to make certain assumptions in our calculation of the power of our study. To reach a power of 80%, and assuming a population variance of 10 and a treatment effect size of 5 minutes, we would require 126 participants, with 63 participants for the control group and 63 participants for the combined treatment groups.

$$k = \frac{n_2}{n_1} = 1$$

$$n_1 = \frac{(\sigma_1^2 + \sigma_2^2/K)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

$$n_1 = \frac{(10^2 + 10^2/1)(1.96 + 0.84)^2}{5^2}$$

$$n_1 = 63$$

$$n_2 = K * n_1 = 63$$

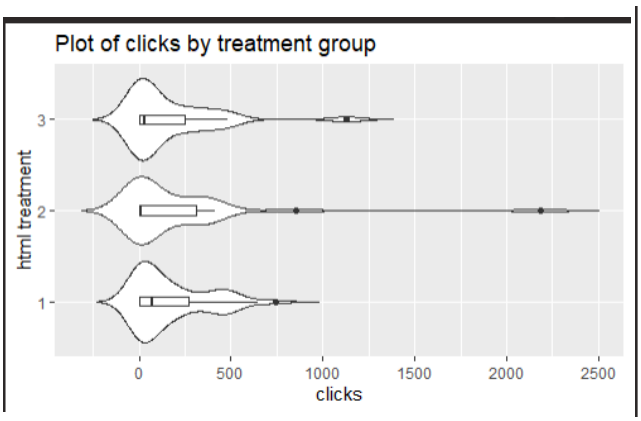
$\Delta = |\mu_2 - \mu_1|$ = absolute difference between two means
σ_1, σ_2 = variance of mean #1 and #2
n_1 = sample size for group #1
n_2 = sample size for group #2
α = probability of type I error (usually 0.05)
β = probability of type II error (usually 0.2)
z = critical Z value for a given α or β
k = ratio of sample size for group #2 to group #1

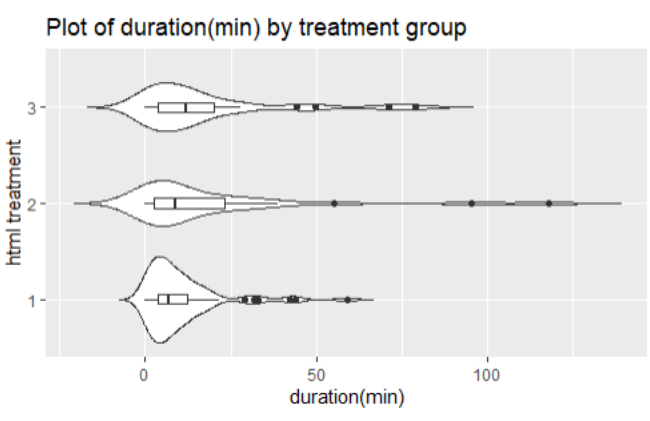
E.v Randomization Process

The experiment utilized randomization through a computer program randomization. Each time the link to the game is clicked, it randomly sends the user to one of the three sites. Users are sent to the control with 50% probability, treatment 1 with 25% probability, and treatment 2 with 25% probability. We were advised by our Professor Micah Redman to split our subjects in this way since there is such a small difference between the two treatment groups. This way if needed we can combine our treatment groups during analysis to gain more power.

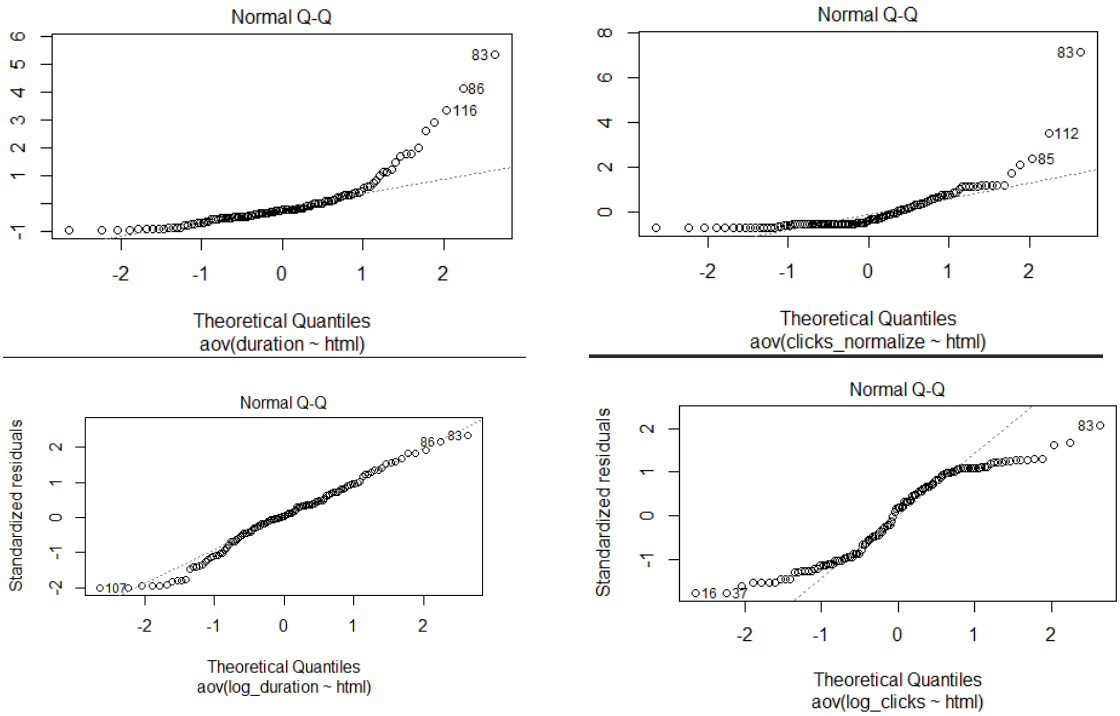
E.vi Observations and Outcome Measures

Our potential outcome variables are clicks and/or duration in minutes. We plot these variables for a quick view of its distribution.





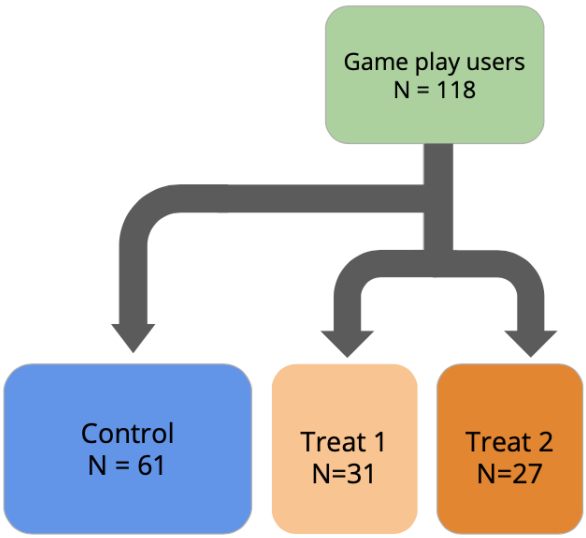
When visualizing the number of clicks and duration (in minutes) we notice a high right skew in the data. In order for us to run linear regression models we check for the normality assumption and find that the data don't fit the regression lines. Thus, we log clicks and duration in order to run tests on linear regression models.



The top QQ plots show lm(duration ~ treatment) and lm(clicks ~ treatment)
The bottom QQ plots show lm(log(duration)~ treatment) and lm(log(clicks) ~ treatment)
The QQ plot for lm(log(clicks) ~ treatment) may not exactly satisfy the normality assumption; however, we will look to use clicks in linear regression for further analysis. We will also test for the statistical significance of the combination of clicks and duration as our engagement outcome variable.

F. Data

Treatment vs Control Flow Diagram



G. Results

```r
res.aov <- aov(log_duration~ html, data = d)
summary(res.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
html          2    1.0  0.4976   0.389  0.678
Residuals   116  148.2  1.2778
```

```r
res.aov <- aov(log_clicks~ html, data = d)
summary(res.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
html          2    7.2   3.591   0.767  0.467
Residuals   116  543.3   4.684
```

Anova tests reveal that the treatment effect has no significance towards measuring log(duration) and log(clicks). We then define engagement to be the combination of the two variables and look to test for statistical significance.

```r
d$engagement =  0.05 * d$log_duration + 0.95 * d$log_clicks
```

```r
res.aov <- aov(engagement~ html, data = d)
summary(res.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
html          2    6.3   3.131   0.715  0.491
Residuals   116  507.9   4.379
```

The treatment effect has no significance towards measuring engagement. We find that solely using log(clicks) instead of engagement provides a better p-value and the p-value for the anova test for engagement has a decrease in p-value if more weight is given to log(clicks) as opposed to log(duration). We would like to note that we are hesitant on selecting the weights that best lower the p-value as this can be a form of p-hacking.

We were also interested in the effect of combining the two treatment groups to have a 1:1 ratio in control to treatment population.

```r
res.aov <- aov(log_clicks~ html, data = d)
summary(res.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
html          1    6.7   6.687   1.439  0.233
Residuals   117  543.8   4.648
```

We test for the statistical significance of the treatment effect on log(clicks) when combining the treatment groups together and find our lowest p-value of 0.233. Combining the treatment groups does, indeed, lower the p-value; however, none of our linear models show statistical significance.

Lastly, to confront our sharp null hypothesis we used randomization inference and our p-values were very similar to that of our Anova tests. log(clicks) was our main variable of interest. The p-value is nowhere near 0.05 and we are not able to reject the sharp null hypothesis.

```{r}
# Randomization Inference
log_clicks_group_mean <- data.table(d)[, .(mean_log_clicks =
mean(log_clicks)), keyby='html']

log_clicks_ate <- log_clicks_group_mean[, diff(mean_log_clicks)]
```

```{r}
n <- nrow(d)

random_ate <- function() {
  random_group_mean <- data.table(d)[, random_treatment:=
sample(c(0,1), replace=TRUE, size=n)]
  random_group_mean <- random_group_mean[, .(mean_log_clicks =
mean(log_clicks)), keyby='random_treatment']
  ate <- random_group_mean[, diff(mean_log_clicks)]
  return(ate)
}

randomization_inference <- function(num_simulations) {
  distribution <- NA
  for (i in 1:num_simulations) {
    distribution[i] <- random_ate()
  }
  return(distribution)
}
```

```{r}
log_clicks_distribution <- randomization_inference(10000)
```

```{r}
p_value <- mean(abs(log_clicks_distribution) > abs(log_clicks_ate))
c(p_value)
```

```
[1] 0.2396
```

H. Conclusions

Our study found that neither reward systems nor prior knowledge of reward systems had any significant effects on player behavior. Before we use our study to make broad inferences regarding reward systems in general, it may be useful to reflect on potential weaknesses of our design. Obviously, with a greater sample size, our study would have seen a corresponding increase in statistical power, and we would be able to measure the outcome effect with greater certainty. Also, if we had some idea of what our effect size or population variance might be, we might have also been able to optimize the parameters of our study further.

Additionally, it is possible that our study design confounded the true effects of reward systems. Our initial reward of a chance to win $25 possibly overshadowed the in-game incentive system of points and font change options, reducing the effect of our implemented reward system. Another possibility is that the limited information we were able to gather was insufficient to fully capture the relationship between treatment and effect; more detailed observations with other metrics, such as perceived difficulty, player age/sex, gaming history, could help create more informed and effective models. We should also note that our experiment does not accurately simulate real-life user behavior in that users typically pick games/services of their own volition; by offering an incentive for players to participate in our experiment (via Amazon giftcards), players who might otherwise not have played this game may have been included in our participants. While randomization should ensure that this affects all groups appropriately, it does not account for the fact that real-life reward systems of games/applications are more likely to affect consumers who were already likely to seek out those specific games/applications, while our experiment is composed of participants who are simply interested in earning $25.

However, if we consider our results to be significant and meaningful, we would conclude that a simple reward system like our implementation is less effective than one might expect. Earning points and cosmetic changes showed little to no effect on participant behavior, and correspondingly, knowing the intent of the reward system made no discernible difference on participant behavior. This demonstrates that this sort of simple extrinsic reward system is less effective than hypothesized, but it should be taken into consideration that many versions and variants of reward systems are implemented in various games and applications. It is entirely possible that more sophisticated reward systems could affect user behavior. Involvement of exposure to other participants (e.g. publicly available high score metrics, social comparison/engagement) could be used as a powerful motivator to increase play/use time, along with more tangible rewards (e.g. monetary rewards for increased use time), could affect user behavior significantly differently than what our experiment has demonstrated.

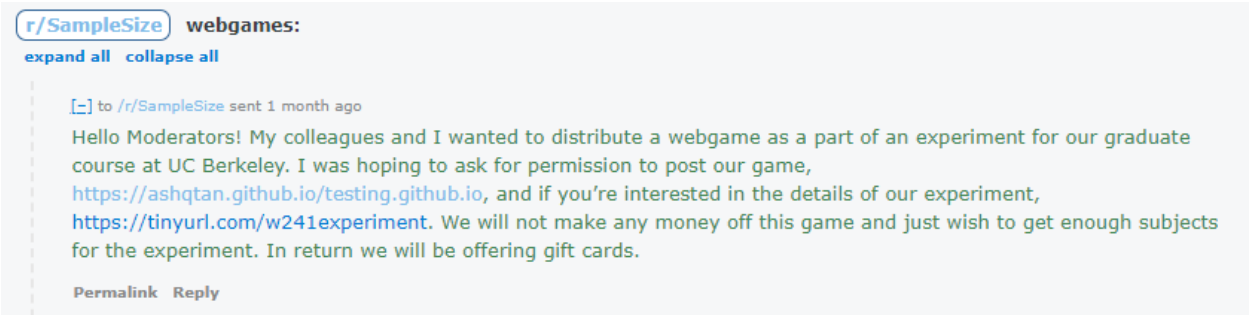
I. Limitations and Future Improvements

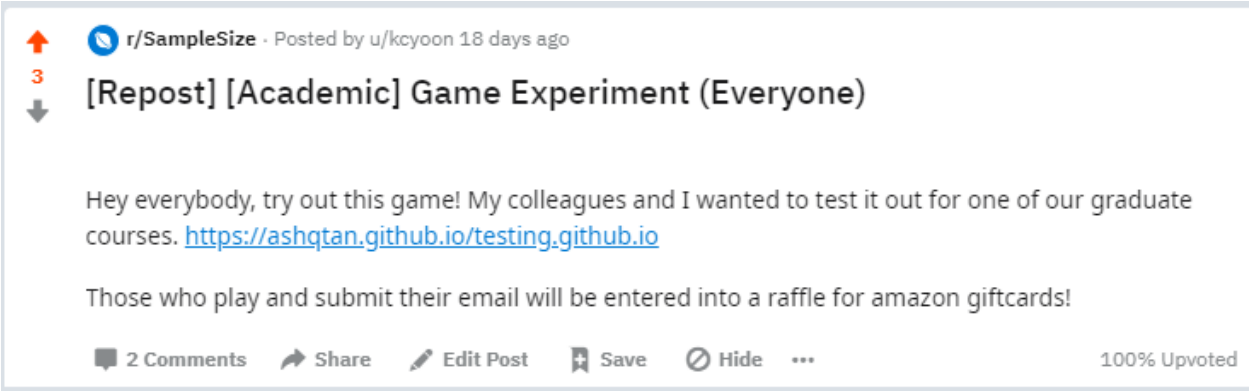
This project was limited by google analytics in various ways. Google Analytics allows for retrieval of precise user activity data on our website htmls. We were able to collect duration and different types of clicking. However, Google Analytics does not track any type of covariates. Because we were not able to track demographics of our users such as age, location, gender we could not use covariates to reduce our standard errors and do a more in depth analysis. Originally we attempted to use our clicks feature as a covariate but notice include this caused a change in sign on the treatment effect. The covariates listed below are in fact no covariates since they depend directly on the treatment. They are outcomes and after seeing this we combined it with the duration outcome.

	Dependent variable:	
	log_duration	
	(1)	(2)
html2	0.146	-0.115
	(0.300)	(0.205)
html3	0.178	-0.130
	(0.314)	(0.217)
clicks_normalize		0.002***
		(0.0003)
done		0.067***
		(0.008)
Constant	1.823***	0.861***
	(0.174)	(0.144)
Observations	119	119
R2	0.004	0.550
Adjusted R2	-0.014	0.534
Residual Std. Error	1.359 (df = 116)	0.922 (df = 114)
F Statistic	0.211 (df = 2; 116)	34.780*** (df = 4; 114)
Note:		*p<0.1; **p<0.05; ***p<0.01

Appendix A - Recruitment Process

(Reddit post, message to moderators)

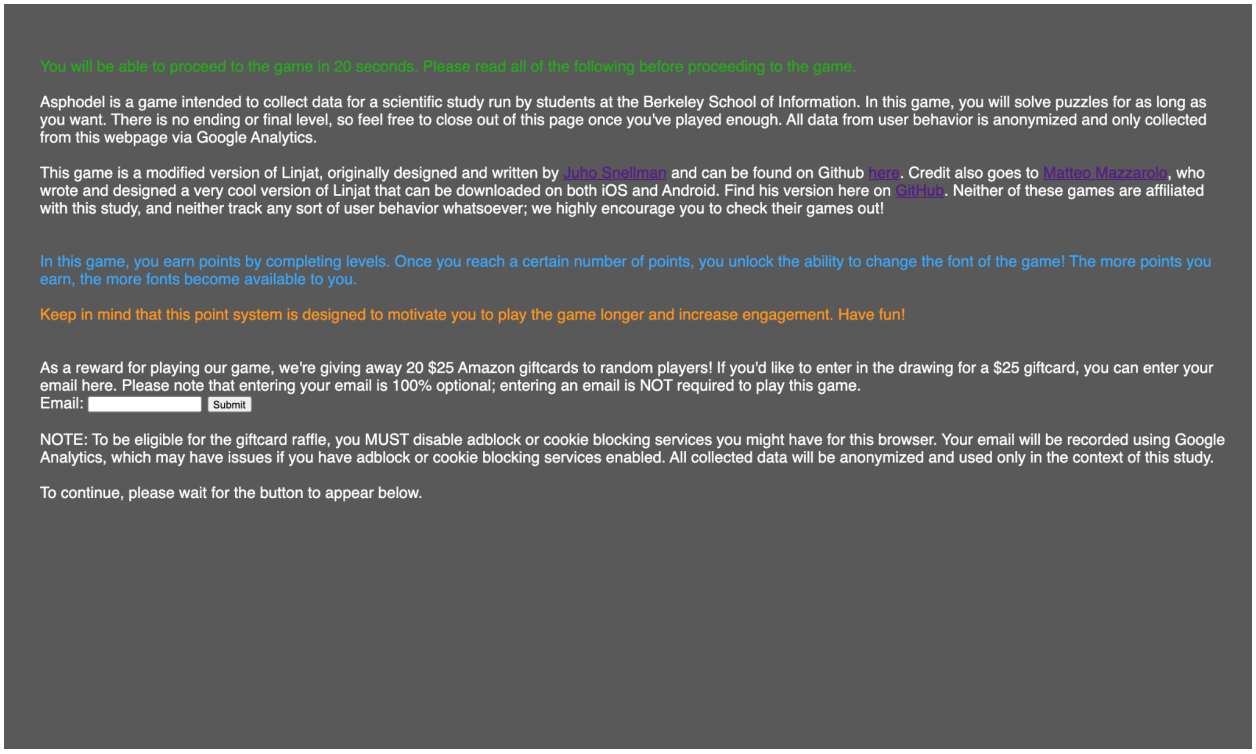




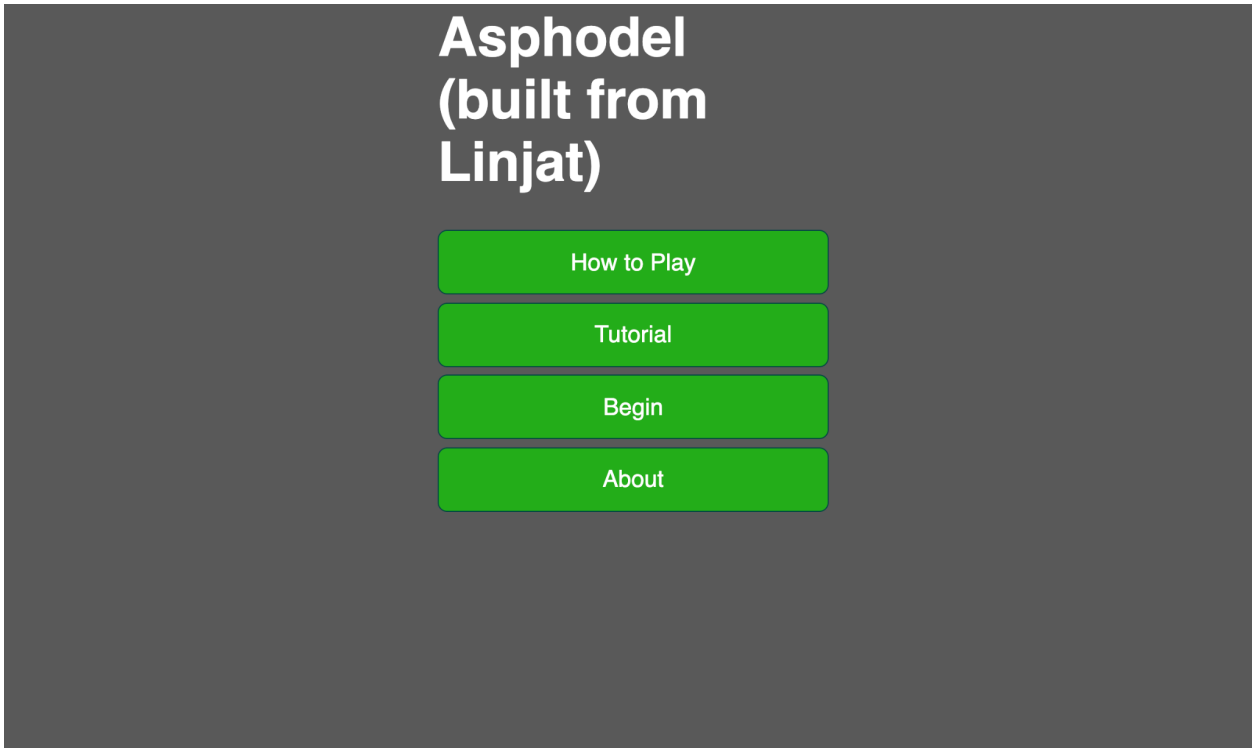
Appendix B - Control/Treatment

(Control group page)

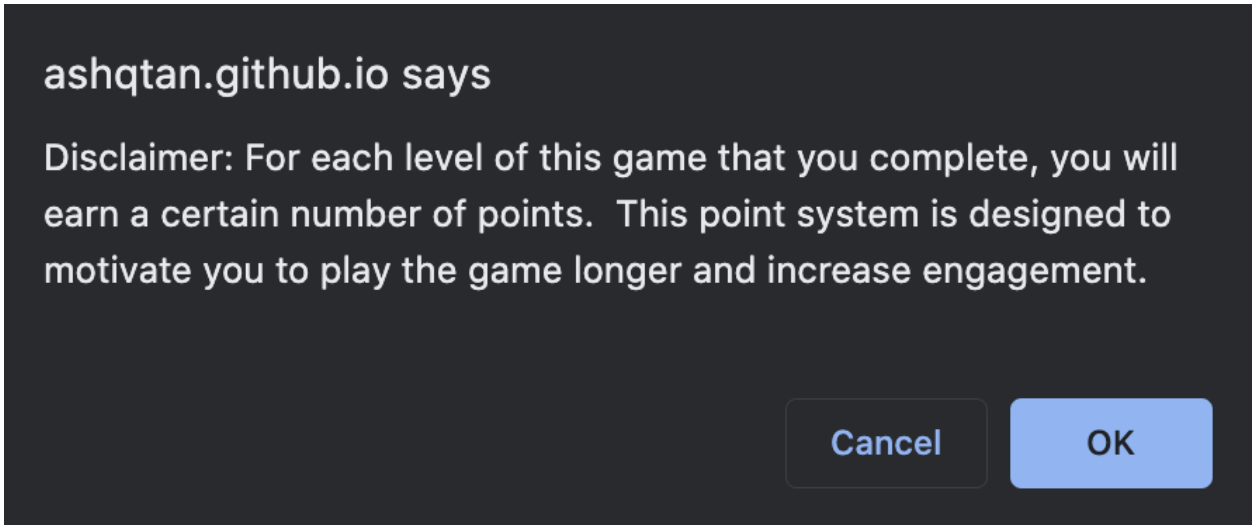
Participants are first directed to the game's landing page:



The text of this page depends on the randomly-selected group of the player: the control group receives only the green and white text (no blue or orange text), the first treatment group receives all text (green, white, blue, orange text), and the second treatment group receives green, blue, and orange text (no orange text). From there, they proceed to the actual game.

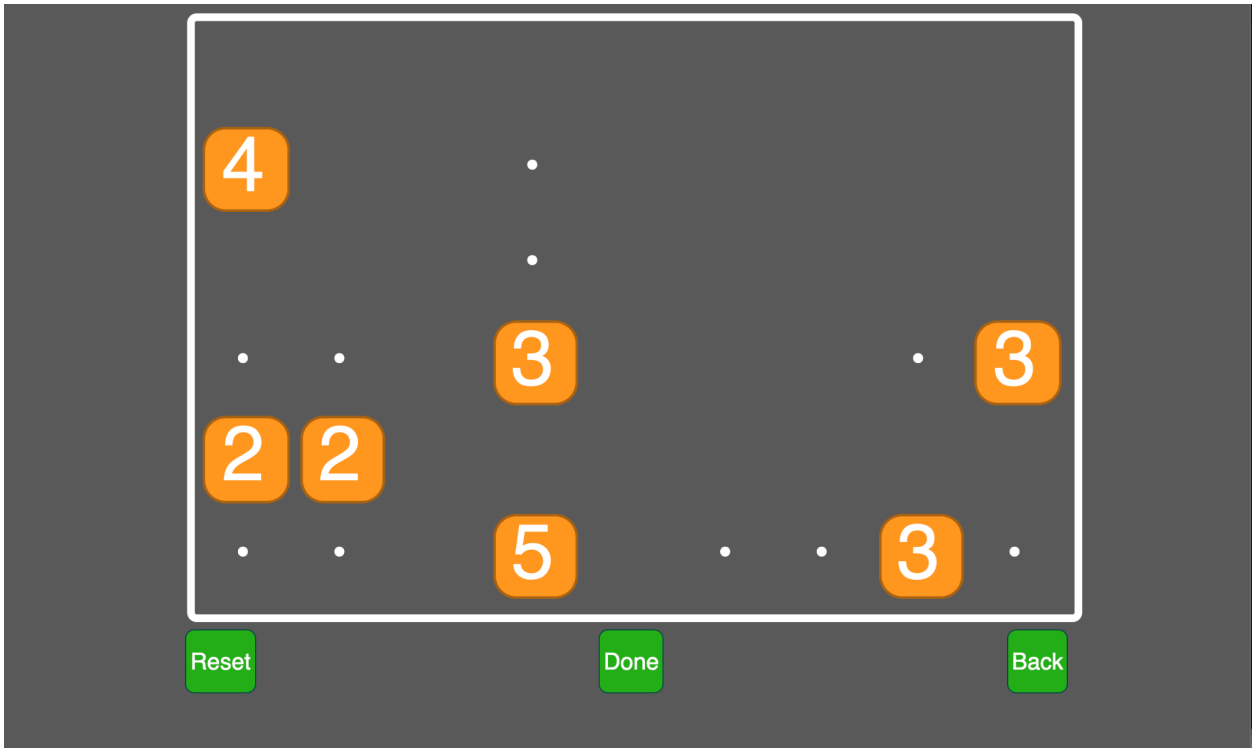


All groups are given the same menu, with options to go through game instructions and tutorial before playing. When clicking on begin, treatment group 1 is given a disclaimer they must acknowledge before proceeding to the actual gameplay.



The gameplay and level design for all groups is exactly the same, with the only difference being UI changes to reflect the reward systems.

Control:



Treatment 1 and 2: