

# The Effects of Location and Color of “Sign in” Button on Online Shopping Platforms

Chun Zhou, Weijia Suo, Yuxuan Mei, Ji Qi, Qianru Ai

12/9/2021

## Introduction

Online shopping platforms have become an essential part of everyone's daily life. Major players in the e-commerce industries include Amazon, Alibaba, Etsy and much more, which “receive valuations two to four times higher... than companies with other business models.” (Libert, Wind, and Fenley 2014). The 2020 COVID 19 drives E-commerce growth even more. Amazon's founder Jeff Bezos was ranked as second richest man (Forbes 2021) in the United States with his online shopping business.

Most online shopping platform firms generate revenue from three sources: pay-per-click search advertising, pay-per-impression display advertising and most importantly, membership fees (Ju-Yeon, Eric and Jisu 2018). In order to encourage user subscription as members and increase user stickiness, online shopping platforms will need to encourage customer sign up and sign in for further services and profits. Membership tier has a positive effect on content generating, review rating and review delay, hence improve customer experience (Dongpu, Yili and Kanliang Wang 2018).

To encourage membership tier and user sign in, shopping website button design is crucial. Studies have found that effects of button design characteristics (i.e., button size, button spacing, visual feedback and button shape) on users' touchscreen performance, mental workload and preference (Da, Juan, and Shuang 2018).

Despite assumed benefits of button design characteristics, little is known about the effect of which type of design will enable users to locate sign in button most efficiently. We were interested in the effect of color and location of sign in button on the speed of shopping websites users to locate it. We believed that the easiness for users to locate the sign in button will increase the membership tier hence better serve the customer and lead to higher member related revenues. The implication of this study can be useful to e-commerce business, consumers, scholars, stakeholders and investors.

To observe this, we created an experiment with 169 randomly generated participants across networks and Amazon Mechanical Turk to test how much time they took to locate the “Sign In” button on the assigned webpage. These participants will participate on our designed website at [https://weijia.io/Experiment/AB-Login\\_Button](https://weijia.io/Experiment/AB-Login_Button) and answer a survey on personal information such as age, gender, educational level and frequency of using online shopping platforms. Then they will enter a randomly assigned webpage with 25% of chance among control, treatment 1,2 and 3. They are given the task to click on sign in button, and the time they took to perform the task will be recorded. For the control group, the “Sign in” button is in the upper left corner of the page. For treatment 1, the “Sign in” button is at the right hand side, together with the user's head-shot. For treatment 2, the “Sign in” button is in the upper left corner of the page, and is highlighted. For treatment 3, the “Sign in” button is at the right hand side, together with the user's head-shot, and is highlighted. Everything except the “Sign in” button is the same for the four groups.

```
library(data.table)
library(fixest)
```

```
library(broom)
library(lfe)
library(tidyverse)
library(knitr)
library(broom)
library(plyr)
library(ggplot2)
library(lattice)
library(modelsummary)
library(kableExtra)
library(pwr)
library(gridExtra)
```

```
data <- fread('cleaned_data_all.csv')
data[,engagementTime := engagementTime/1000]
kable(data[1:5]) %>%
  kable_styling(latex_options="scale_down") %>% kable_styling(latex_options="hold_position")
```

id	device	country	Age group	Gender	Highest degree	Frequency	title	engagementTime
0.0441247	Desktop	us	55 ~ 64 years old	Female	Bachelor	At least once a week	A	26.374
0.4305218	Desktop	in	25 ~ 34 years old	Male	Bachelor	At least once a month	B	8.326
0.5306845	Desktop	in	35 ~ 44 years old	Female	Bachelor	At least once a day	C	15.180
0.5489083	Desktop	us	35 ~ 44 years old	Male	Bachelor	At least once a week	C	16.591
0.6866643	Desktop	us	25 ~ 34 years old	Male	Bachelor	At least once a day	C	8.753

## Method

### Exploratory Data Analysis

Our Exploratory Data Analysis shows that on average, the control group's engagement time (time spent by a participant to click the sign in button) is slightly higher than the three treatment arms. For our participants demographic information, we found little/no correlation between shopping frequency and engagement time. Surprisingly, Masters and Bachelors are most related with a high engagement time. While seemingly people of age between 55 to 64 are more related with a high engagement time, after further exploration we found that there are only 4 participants in that age group, therefore we don't have enough statistical power to conclude a positive correlation.

```
df_group <- data[, mean(engagementTime), by=title]
df_age <- data[, mean(engagementTime), by='Age group']
df_gender <- data[, mean(engagementTime), by='Gender']
df_edu <- data[, mean(engagementTime), by='Highest degree']
df_freq <- data[, mean(engagementTime), by=Frequency]

y_by_group <- ggplot(df_group, aes(x=title, y=V1)) +
  geom_bar(stat="identity") + labs(x= 'Group', y = 'Engagement time (s)')

y_by_age <- ggplot(df_age, aes(x= reorder(df_age$'Age group',-V1), y=V1)) +
  geom_bar(stat="identity") + labs(x= "Age group", y = 'Engagement time (s)') +
  scale_x_discrete(labels = function(age) str_wrap(age, width=7)) +
  theme(axis.text = element_text(size = 6))
```

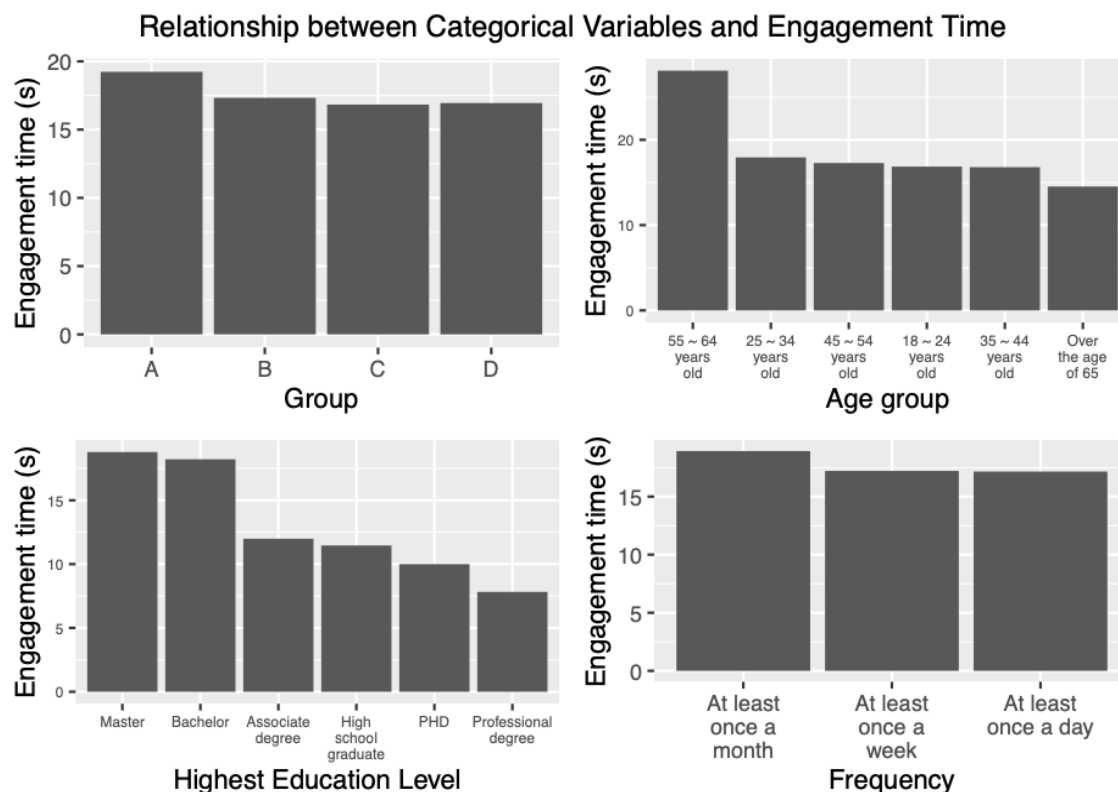
```

y_by_edu <- ggplot(df_edu, aes(x= reorder(df_edu$`Highest degree`, -V1), y=V1)) +
  geom_bar(stat="identity") + labs(x = 'Highest Education Level', y = 'Engagement time (s)') +
  scale_x_discrete(labels = function(`Highest degree`) str_wrap(`Highest degree`, width=2)) +
  theme(axis.text = element_text(size = 6))

y_by_freq <- ggplot(data= df_freq, aes(x= reorder(Frequency, -V1), y=V1))+
  geom_bar(stat="identity") + labs(x = 'Frequency', y = 'Engagement time (s)') +
  scale_x_discrete(labels = function(Frequency) str_wrap(Frequency, width=10))

grid.arrange(grobs = list(y_by_group, y_by_age, y_by_edu, y_by_freq), nrow=2,
  top = "Relationship between Categorical Variables and Engagement Time")

```



## Participants and Randomization

Initially participants come from researchers' personal network. We posted experiments on alumni platforms from previous education, online groups of various interests and locations across the globe. We noticed that participants were lack of incentives to finish the task and the network was not absolutely random. Then we turned to use Amazon Mechanical Turk which has a larger participants pool with various characteristics. We ended up having participants from 6 countries, aging from 18 years old to over the age of 65 and educational background from high school to PhD.

Although there are a few biases that emerge when recruiting only people we personally know, Amazon Mechanical Turk data constitutes over 50% of data set. If with more research funding allowance, we could obtain larger data size from Mechanical Turk for better results.

## Balance Check

We used a two-step approach for our balance check. First of all, we used prop test on all the groups to check the randomization probabilities. Our result showed that treatment arm 2 has a p-value less than 0.05, so we reject the null hypothesis that the true probability is equal to 25%. To find out why this happened, we went back to our randomization code, as well as all survey data we collected. We found that our randomization code was running properly, and the problem was caused by a large number of uncompleted experiments assigned to treatment arm 2.

From there, we also conducted balance checks on pre-experimentation characteristics. Due to limited samples we have and the categorical nature of the characteristics, we didn't directly compare 4 groups we have. Instead, we looked at characteristics that is related with high engagement time (Bachelors, Masters, Age group 55-64), and checked if our groups contained an abnormal number of samples with these characteristics. And our result shows that except for treatment arm 2 (group C), we have similar number of samples in these categories. While it's also worth noticing that aside of Group C, we also have fewer Masters in Group A, which could contribute to improper randomization to some extent.

```
## prop test
resultA <- prop.test(data[title == 'A', .N], data[, .N], p=.25)
resultB <- prop.test(data[title == 'B', .N], data[, .N], p=.25)
resultC <- prop.test(data[title == 'C', .N], data[, .N], p=.25)
resultD <- prop.test(data[title == 'D', .N], data[, .N], p=.25)

balance_check <- data.table(Title = c("A", "B", "C", "D"),
                             `P-value` = c(resultA$p.value, resultB$p.value,
                                             resultC$p.value, resultD$p.value))

kable(balance_check, digits = c(0,3,3,3))
```

Title	P-value
A	0.756
B	0.143
C	0.037
D	0.351

```
## characteristic test
bechelor <- data[data$'Highest degree' == 'Bachelor', .N, title]
setnames(bechelor, 'N', 'Bechelors')
master <- data[data$'Highest degree' == 'Master', .N, title]
setnames(master, 'N', 'Masters')
age_55 <- data[data$'Age group' == '55 ~ 64 years old', .N, title]
setnames(age_55, 'N', 'Age_55_64')
balance_check2 <- merge.data.table(merge.data.table(bechelor, master), age_55, all = TRUE)
balance_check2[is.na(balance_check2)] <- 0
kable(balance_check2, digits = c(0,3,3,3))
```

title	Bechelors	Masters	Age_55_64
A	29	8	1
B	33	12	2
C	18	7	1
D	29	15	0

## Pre-experiment Power Calculation

According to our calculation, in order to detect an effect size that's 30% of the outcome's standard error at a significance level of 0.05 with a power of 0.8, we need a sample size of at least 175. Therefore, we aim to

collect a sample size around 200 to make sure our experiment has enough statistical power.

```
# data[, SD(engagementTime)]
power <- pwr.t.test(n=NULL, d=0.3, sig.level=.05, power=.8)
power <- data.table(` ` = c("d", "Significant Level", "Power", "Required Number of Samples"),
                    Value = c(power$d, power$sig.level, power$power, power$n))
kable(power, digits = c(0,3,3,3))
```

	Value
d	0.300
Significant Level	0.050
Power	0.800
Required Number of Samples	175.385

## Procedure

We first designed the website constituted of two parts: google form to collect relevant personal data and click sign in button experiment. Upon users' landing at the website, they are randomly assigned with a number between 0 and 1 as their unique user ID for us to track the results at the back end of website. Since we have one control group and three treatment arms, the probability for the user to be assigned into one category is 25%. This random assignment happens before all the procedure of filling the Google form and the experiment. Then, participants are asked to fill in four multiple choice questions regarding to age, gender, educational level and frequency of using shopping websites. Finishing on the Google form, participants will be instructed to move on to our virtual shopping website and click on the sign in button. Once they click on the sign in button, the time duration from landing on the shopping website to correctly clicking the button will be recorded by the back end. Finally, participants will be directed to a Thank You page after they successfully finished the clicking button experiment.

## Average Treatment Effects

Based on our data set, we will run a simple regression to investigate the main effect of three different treatments (changing location, changing color, changing both color and location) on the time spent on locating the "Sign In" button on the assigned webpage. Thus, Below are the results of our first regression.

```
ATE <- feols(log(`engagementTime`) ~ title, data = data, se = 'hetero')
ATE %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE", "T-Stat", "P-Value"),
                        digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	2.539	0.144	17.679	0.000
titleB	0.008	0.180	0.043	0.966
titleC	-0.044	0.210	-0.212	0.832
titleD	-0.005	0.189	-0.029	0.977

From the regression table, on average, the time spent on searching the "Sign In" button in treatment 1 was increased by 0.778% compared to those in the control group. This indicates that changing location to the right-hand side, together with the user's head-shot may create some noise for people to locate the correct location of the target. However, the mean value of time used on seeking the the "Sign In" button in treatment 2 and 3 was decreased by 4.45% and 0.539%, respectively. Our result implies that highlighting the color of the target and both highlighting the color and placing the target at the right hand side of the webpage are user-friendly approaches. Especially, highlighting the color of the "Sign In" button could save the users' time by 4.45% compared to the control group, which is more than the treatment 3. However, the



p-values of all three treatments are more than 0.8 and not statistically significant at 0.05, which means that we are not confident to conclude that on average, all three treatment would have effects on the time spent on locating the “Sign In” button. The possible reason is that the small sample size about 169 data points reduces the chance of detecting a true effect.

## Covariates

According to the results from EDA, we found that education level and age are correlated to the engagement time on locating the Sign-in Button. Therefore, we added these two variables in to our simple regression model to check if we can improve the estimates precision. The regression table in below indicates that the average treatment effects of all three arms are close to the those in our first regression model and are also not significant different at 0.05 significant level. Furthermore, the standard errors of those three treatment estimates increases a little bit, which implies that the newly added covariates couldn't increase the precision of our results. Thus, we decided not to add covariates in our model.

```
ATE2 <- feols(log(`engagementTime`) ~ title + `Highest degree` + `Age group`,
              data = data, se = 'hetero')
ATE2 %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE",
                                       "T-Stat", "P-Value"),
                        digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	2.085	0.336	6.200	0.000
titleB	0.056	0.193	0.289	0.773
titleC	0.011	0.224	0.050	0.960
titleD	-0.004	0.189	-0.021	0.984
'Highest degree'Bachelor	0.287	0.277	1.037	0.302
'Highest degree'High school graduate	-0.222	0.456	-0.486	0.628
'Highest degree'Master	0.427	0.303	1.408	0.161
'Highest degree'PHD	0.053	0.331	0.160	0.873
'Highest degree'Professional degree	-0.231	0.370	-0.624	0.534
'Age group'25 ~ 34 years old	0.165	0.189	0.871	0.385
'Age group'35 ~ 44 years old	0.173	0.208	0.832	0.407
'Age group'45 ~ 54 years old	0.316	0.242	1.305	0.194
'Age group'55 ~ 64 years old	0.349	0.657	0.531	0.596
'Age group'Over the age of 65	0.206	0.311	0.662	0.509

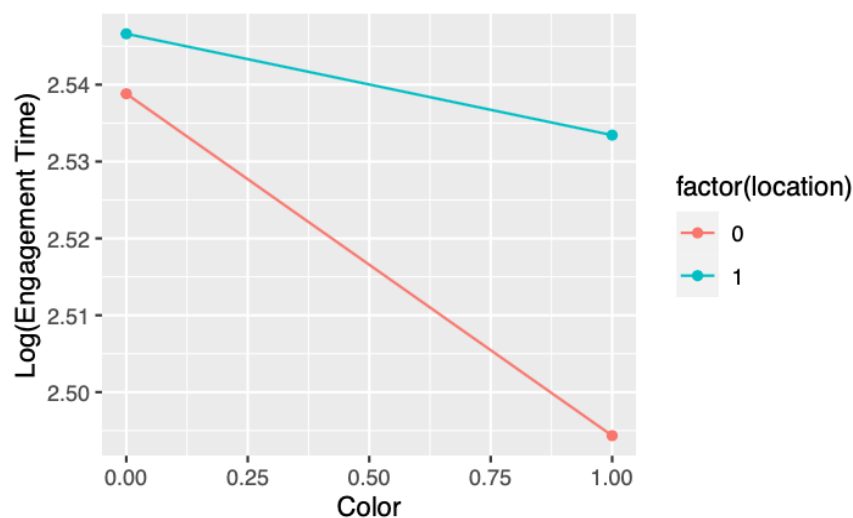
## Factorial Experiment

From another point of view, we can run the regression model for a simple 2 X 2 Factorial Design. So, we create a factor called 'color' (1: highlighting the “Sign In” button vs.0: Not highlighting the “Sign In” button) and another factor named 'location'(1:placing the “Sign In” button on the right-hand side vs. 0: at the upper left corner of the web page). Thus, we will investigate the two main effects and one interaction effect of 'color' and 'location'.

Firstly, we plotted the means to gain a better understanding of the effects that two factors have on the time spent on searching the “Sign In” button. The graph in below shows that two lines are roughly parallel, which means there is highly possible no interaction effect between 'color' and 'location'. In other words, the effect of 'color' has on the time spent on locating the target doesn't depends on 'location'. In order to inspect the effects that each factor on the dependent variable, we run the regression of color and location on the time used for searching the location of the “Sign In” button.

```
data[(title == 'C') | (title == 'D'), color := 1]
data[(title != 'C') & (title != 'D'), color := 0]
data[(title == 'B') | (title == 'D'), location := 1]
data[(title != 'B') & (title != 'D'), location := 0]
```

```
data2 <- data[, mean(log(engagementTime)), by = .(`color`, `location`)]
data2 %>% ggplot() + aes(x = color, y = V1, color = factor(location)) +
  geom_line(aes(group = location)) +
  labs(x = 'Color', y = 'Log(Engagement Time)') +
  geom_point()
```



```
FD <- feols(log(`engagementTime`) ~ color * location, data = data, se = 'hetero')
FD %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE",
                                     "T-Stat", "P-Value"),
                      digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	2.539	0.144	17.679	0.000
color	-0.044	0.210	-0.212	0.832
location	0.008	0.180	0.043	0.966
color:location	0.031	0.266	0.118	0.906

According to the regression table, we can notice that the coefficients of color (-0.0445) and location (0.0078) are the same as the coefficients of treatment 2 (-0.0445) and 1 (0.0078), respectively. In addition, the coefficient of treatment 3 (-0.005389) is equal to the coefficients of color (-0.0445) and location (0.0078) plus the coefficient of the interaction of two factors (0.0313). This indicates that the effect of both highlighting the color and placing the target at the right hand side of the webpage is equal to the sum of effects of color, location and the interaction of color and location. It worth noting that p-values associated with color (0.832), location (0.966) and the interaction of color and location (0.906) are all more than 0.05, which means two factors and one interaction term didn't have statistically significant effects on the time spent on locating the target. The conclusion is the same as the first simple regression we run before.

## Post-experiment Power Analysis

We realized that our outcome variable (engagement time) has a large standard deviation, which is negatively impacting our experimentation power. In order to detect an effect size around 1 second, we need our Cohen's  $d$  to be around 0.05. According to our calculation, in order to detect an effect size that's 5% of the outcome's standard error at a significance level of 0.05 with a power of 0.8, we need a sample size of at least 6300. Therefore, in future experiments, we need a much larger sample size to make sure our experiment has enough statistical power to make up for the large standard deviation in our outcome variable.

```
data[, SD(engagementTime)]
```

```
## [1] 16.03719
```

```
power2 <- pwr.t.test(n=NULL, d=0.05, sig.level=.05, power=.8)
power2 <- data.table(` ` = c("d", "Significant Level", "Power", "Required Number of Samples"),
                      Value = c(power2$d, power2$sig.level, power2$power, power2$n))
kable(power2)
```

	Value
d	0.050
Significant Level	0.050
Power	0.800
Required Number of Samples	6280.049

## Limitations

To see if using cell phones would make any difference, we calculate the CATE of device.

```
data[, 'any_treatment'] <- data[, 'title'] != 'A'
reg_phone <- lm(log(engagementTime) ~ any_treatment * device, data=data)
reg_phone %>% tidy() %>% kable(col.names = c("Predictor", "Coefficient", "SE", "T-Stat",
                                           "P-Value"), digits = c(0, 3, 3, 3, 3), align = 'c')
```

Predictor	Coefficient	SE	T-Stat	P-Value
(Intercept)	2.566	0.137	18.728	0.000
any_treatmentTRUE	-0.008	0.157	-0.054	0.957
devicePhone	-0.357	0.500	-0.714	0.476
any_treatmentTRUE:devicePhone	-0.088	0.585	-0.150	0.881

Then the CATE =  $-0.008 - 0.088 = -0.096$ , approximately 0. However, we should notice that there are differences exists between using computers and phones. Firstly, our experiments were based on computer-sized web pages, and we did not design web pages specifically for cell phones. However, we were unable to prevent interviewees from using their cell phones for the experiment. If the interviewees use their cell phones, the web pages they see will be small and they will need to manually zoom in. This may have an impact on the results of the experiment. The good thing is that our dataset shows that only a small fraction of the users used cell phones for the experiment. So even if there is an effect, it will be small. Besides, we use the time stamp of the user's end page minus the time stamp of the start page for the engagement time calculation. However, we found that some cellphone users' end page time is not correctly recorded. Therefore, the engagement time of mobile users have some errors.

There are other limitations exist. For PC users, We have received some feedback that Google Chrome would alert users to the safety of our website. Unfortunately, we could not fix it, so this may have some impact on our results. Also, factors such as did users drink coffee or not, their health condition, and network environment may also have an effect on the results.



## Conclusion

We did not find there is a significant difference across control and tree treatment arms, so we assume that the color and location of “Sign in” button can not effect the users’ experience. We think there are two reasons. Firstly, we have a small sample size. According to the power test, we need at least 1571 samples to make this experiment have power, but we only have 169 samples. Secondly, the outcome has a high standard deviation, because we did not find any useful covariate in our dataset.

For further exploration, we have some thoughts. Firstly, we should collect more data to have enough power, and we could collect more information on factors that may have impacts on the engagement time. Then, we could create new treatment arms based on factors mentioned in papers about users’ web browsing habits. Lastly, as the factorial analysis has the same result, we can actually create fewer arms. Thus, the number of samples in each group would increase.

## Bibliography

- [1]Libert, Barry, Yoram (Jerry) Wind and Megan Beck Fenley (2014), *What Airbnb,Uber, and Alibaba Have in Common*. Available at: <<https://hbr.org/2014/11/what-airbnb-uber-and-alibaba-have-in-common/>> [Accessed 2 December 2021].
- [2]Ju-Yeon, Eric, Jisu and Robert W. Palmatier 2018, *The Effect of Online Shopping Platform Strategies on Search, Display and Membership Revenues*, Journal of Retailing 94 (3,2018) 247-264.
- [3]Dongpu, Yili, and Kanliang et.al, *Effects of Membership Tier on User Content Generation Behaviors: Evidence from Online Reviews*, Electron Commer Res (2018) 18:457-483. Available at: <<https://doi.org/10.1007/s10660-017-9266-7/>> [Accessed 3 December 2021].
- [4]Da, Juan, and Shuang et.al 2018, *Effects of Button Design Characteristics on Performance and Perceptions of Touch Screen Use*, International Journal of Industrial Ergonomics 64 (2018) 59-68.

## Appendix

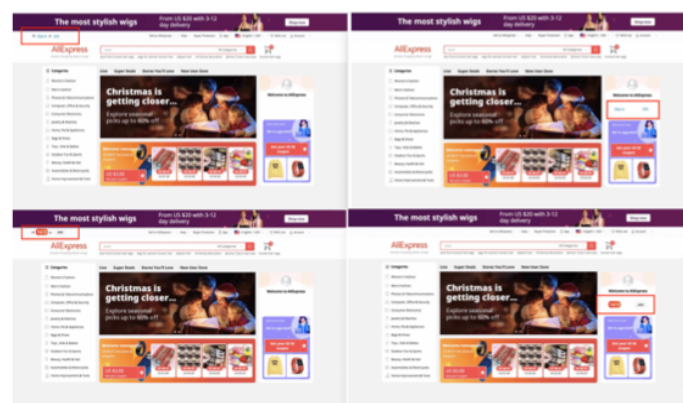


Figure 1: Control & Arms

Sequentially, left-to-right then top-to-bottom are websites for control, treatment 1, treatment2, and treatment 3. The red boxes are only used to mark the location of the buttons in the diagram and are not present in the experiment pages.

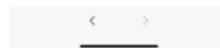


Figure 2: Page on Cellphone

### Questionnaire

Please complete the following questions to help us obtain more accurate results .  
Your unique id is: 0.661216368867614

#### About you

qianrui@bu.edu (not shared) [Switch account](#)

\* Required

What is your unique id in this page (the red number)? \*

Your answer

What's your age group? \*

☐ Under the age of 18  
☐ 18 ~ 24 years old  
☐ 25 ~ 34 years old  
☐ 35 ~ 44 years old  
☐ 45 ~ 54 years old  
☐ 55 ~ 64 years old  
☐ Over the age of 65

What is your gender? \*

☐ Male  
☐ Female

What is the highest degree or level of school you have completed? \*

☐ No schooling completed  
☐ Some high school, no diploma  
☐ High school graduate  
☐ Associate degree  
☐ Bachelor's degree  
☐ Master's degree  
☐ Professional degree  
☐ Doctorate degree

How often do you use shopping websites? \*

☐ At least once a day  
☐ At least once a week  
☐ At least once a month  
☐ Never

Notice! This is not the end of experiment. You must click the below orange button ("Click to start") after click submit button to start the experiment.

Submit
Clear form

Click to start

Figure 3: Questionnaire