Analyzing Impact of Racial Similarity on Movie Rating

Team 12

2/29/2020

Contents

Introduction

One of the biggest challenges faced by streaming services like Netflix is to make users consume new content. With more and more media houses coming up with their own content consumption platforms, popular Shows like The Office, Friends, Marvel's movie catalog are going off Netflix. The platform still has a strong collection of original content like House of Cards and the Black mirror, but it is becoming harder and harder for them to make users stay without the pull of traditional titles from mega studios. On top of this, because of the availability of such a cast array of content, showing the right content at the right time to grab a user's attention is of utmost importance.

One way to grab attention is by putting up interesting and compelling visual cues to its users. Netflix having different types of people on its platform, it is important to curate this content as per preferences to get maximum eyeballs and clicks towards new content it wants users to consume. Netflix utilizes users' behavior data on its website to conduct tests and determine the best content to show you at any point. One such test is to change the thumbnail of the show to make users click on it.

Our team was curious about how users racial origins affected their perception of a certain show/movie on Netflix. MSBA being a mix of Students from India, China, and the US, it was the perfect opportunity to conduct an experiment within the cohort to determine if this hypothesis was true.

Experiment Design

Through this experiment, the question we wanted to answer is, 'Are Users more likely to click on thumbnails that feature a person from their own race?' namely, are Indians more inclined to click on a thumbnail if they see an Indian person in the thumbnail? If the hypothesis was validated, then this information could be utilized to get more clicks and views for unknown shows and also studios could strategically cast actors from different racial backgrounds to target it towards a specific audience.

To remove any bias from the experiment we designed the experiment as follows: 1. Each user will be shown 5 thumbnails from Netflix which they have to rank in order of preference. The ranking system is better than providing an absolute score for the images as it is easier to compare than to absolutely judge something. 2. All 5 thumbnails would be from the same genre, to control for users preference. 3. Out of the 5 thumbnails, 4 would contain American Actors and one thumbnail would contain the Actor from the same origin as the user.

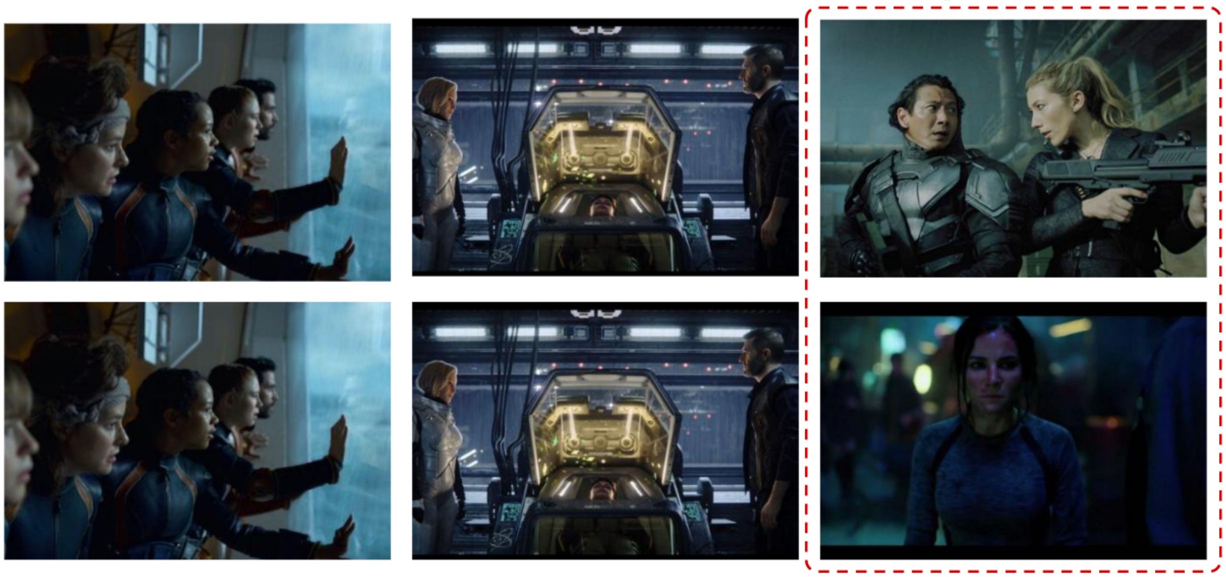


Figure 1: The above example indicates a sample survey: the first two images are the same for both test and control groups whereas the third image has an Asian character for test

4. We would then collect and analyze the results of the survey to determine if people prefer shows having actors of their own origin.

The experiment would be sent to students in the form of a Qualtrics Survey, provided by the University of Minnesota.

Threats to Establishing Causality

We need to validate the design of the experiment and assess the potential threats to the causal claims we hope to make using the controlled experiment. This is vital because some of the threats can directly hamper the causal relationship between ad spends and the response variable, and only in a setting where the assumptions hold good, and the threats are benign, we can proceed to attribute causality.
1. A comparison of the treatment and control groups on observed features tells us there aren't any statistically significant differences between the two groups. We check this using a Randomization Test 2. Statistical Power: We need to have a required number of samples to be reasonably certain that we can measure a minimum difference between the two groups. We check for this using the Power Test 3. Interference - Since the survey was sent to the whole cohort at the same time, and answered within the day, we can assume there was no interference between those taking it. Also, the Respondents were unaware of if they were in Control or Test group and what was being tested in the survey 4. Valid Control group - The control groups get exactly the same images as a treatment except for one, which acts reasonably as a valid control against ads aimed at making the consumer buy the subscription. 5. Social Desirability Bias: There is no Social desirability bias here, as the respondents are unaware of the test and control and test groups are not aware of each other's surveys 6. Demand Effects: This would not be an issue here as the survey is just ranking of shows 7. The Oprah Effect: Again, This factor will not be an issue as the shows are relatively unknown and the experiment is being conducted on the same day. 8. Measurement Error: In this experiment we are comparing the relative rankings of movies w.r.t each other. Since all images will have an avbsolute rank assigned to it, there should not be any measurement error

A Few Considerations

There were a few things the team had to account for when designing the experiment.

Experiment Audience

Since the MSBA cohort consists of 92 students of Indian and Chinese origin out of 117, we only considered testing on these two as we would not have sufficient sample sizes to test. Since these two classes were relatively smaller in size and imbalances for traditional randomization, we resorted to block-randomization to ensure test and control groups are homogeneous. The groups were divided across the following factors: + Race(Indian/Chinese) + Gender + Age + Background (Technical/Non-Technical)

After block randomization, there were 43 students in the treatment group and 44 in the control group. The team was expecting 87 * 6 = 522 responses from the cohort to test the hypothesis.

Popularity Checks

Shows like Stranger Things, House of Cards and Black Mirror are widely known and are watched by audiences of all ages and origins. If these 'famous' shows were included in the survey, there was a high chance the results would be skewed in their favor just because of popularity. To control for this, the team decided to only include shows which were relatively unknown to the target audience. There was no specific selection metric for these shows, they were solely selected on the designer's knowledge of the audience and their preferences.

A similar approach was taken while selecting cast i.e. famous actors and actresses were not selected in the thumbnail as they had the potential to bias the results

Number of Responses

We were aware that not all the responses we get from students would be valid. Some would not choose to answer, some would take the incorrect survey, etc. Apart from this, we were aware that there are only about 100 students in the cohort who were eligible to take the survey, so we decided to include 5 sets of thumbnails of equal genres in each of the tests as well as control surveys. This would solve 2 issues: + There would be 6 responses from each student significantly increasing the number of collected responses which would allow us to conclude results with more statistical certainty. + The 5 questions in the survey would be from the genres: Romance, Sci-Fi, Teen, Drama, Thriller, and Comedy which could also control for a user's preferences across genres.S

```r
suppressPackageStartupMessages({
  library(dplyr)
  library(ggplot2)
  library(MESS)
  library(jtools)
  library(pander)
})
```

```r
power_t_test(n=NULL,type=c("two.sample"), alternative="two.sided",
             delta=0.5,power=0.8,sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76576
##           delta = 0.5
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

From the above power t-test, we find that the optimal number of users to calculate a difference of 1 with 80% power is 63. We believe this is an achievable number of responses and we will be able to conclusively establish causality in this experiment.

Experiment Results

Data Cleaning

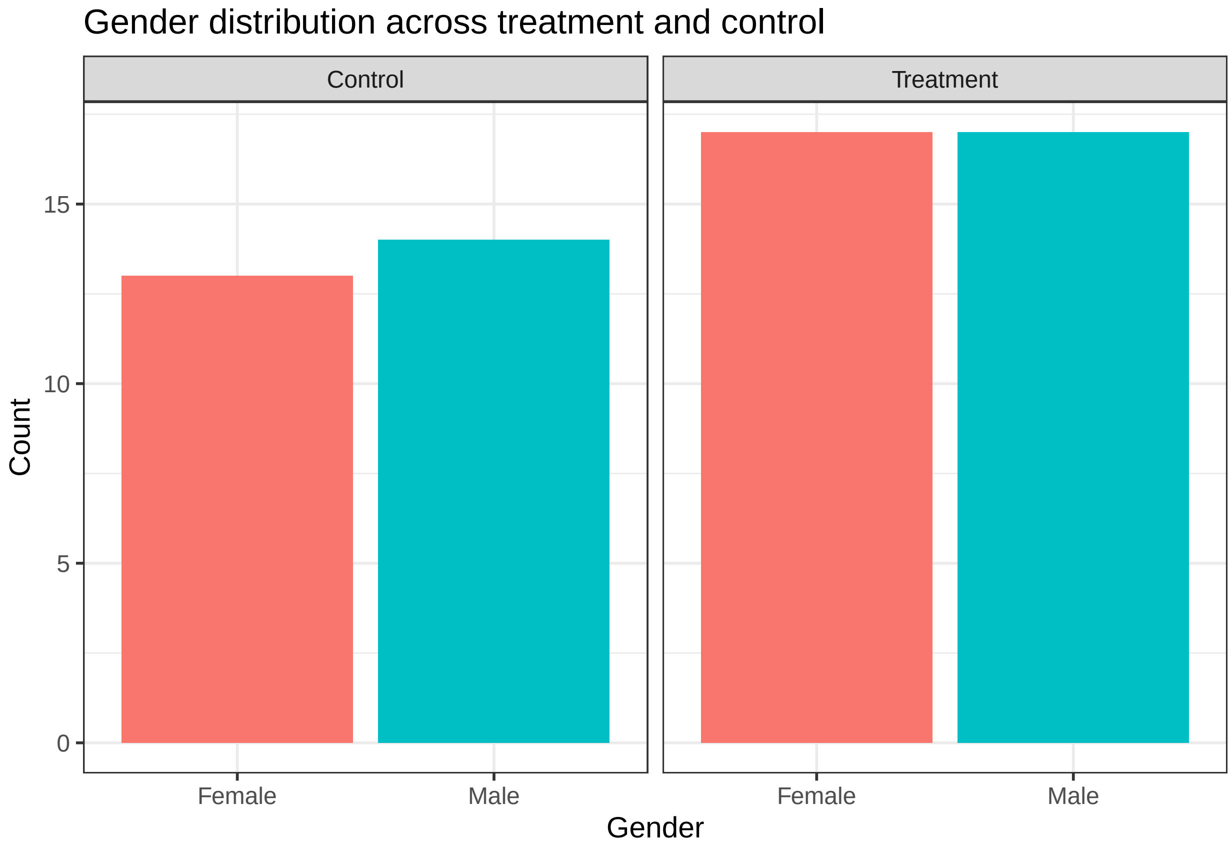
The data obtained from the survey needed cleaning as there were a lot of errors from students filling up the survey. These inconsistent instances had to be filtered out before carrying out the analysis. The issues included: 1. Selecting the wrong treatment group (Indian/Chinese) 2. Not changing the order and leaving the responses as is 3. Not answering all the questions in the survey 4. Completing the survey in <30 seconds (clearly not enough to complete the survey)

The team expected around close to 480 results but because of these inconsistencies, they had to settle for 393 clean survey results.

Data Exploration

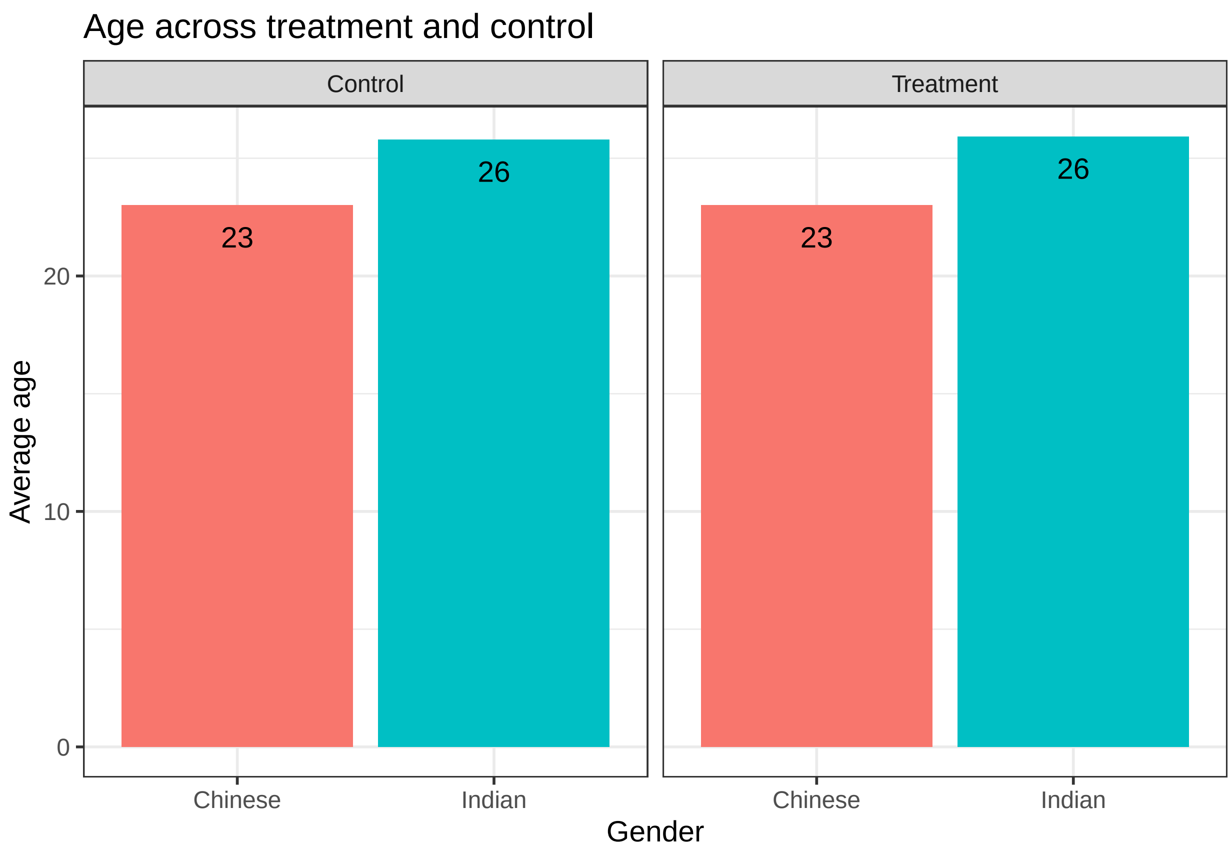
The team had 393 valid survey responses at hand. Out of which 230 were from Male students and 163 were from female students.

```r
setwd('C:\\Users\\vengu002\\Downloads\\')
data = read.csv('surveyCleaned.csv')
test_control = read.csv('test_control.txt', sep = '\t')

data_final = data %>%
  left_join(test_control, by = 'Name') %>%
  mutate(up_flag = ifelse(Ranking > originalRank, 1, 0)) %>%
  filter(is.na(Flag) == FALSE)

data_final %>% select(Name, Gender, Race, Flag) %>%
  unique() %>%
  mutate(Gender = ifelse(Gender == 'M', 'Male', 'Female')) %>%
  ggplot(aes(x = Gender, fill = Gender)) +
  geom_bar() +
  facet_grid(cols = vars(Flag)) +
  labs(title = 'Gender distribution across treatment and control',
       y = 'Count', x = 'Gender') +
  theme_bw() +
  theme(legend.position = "none")
```



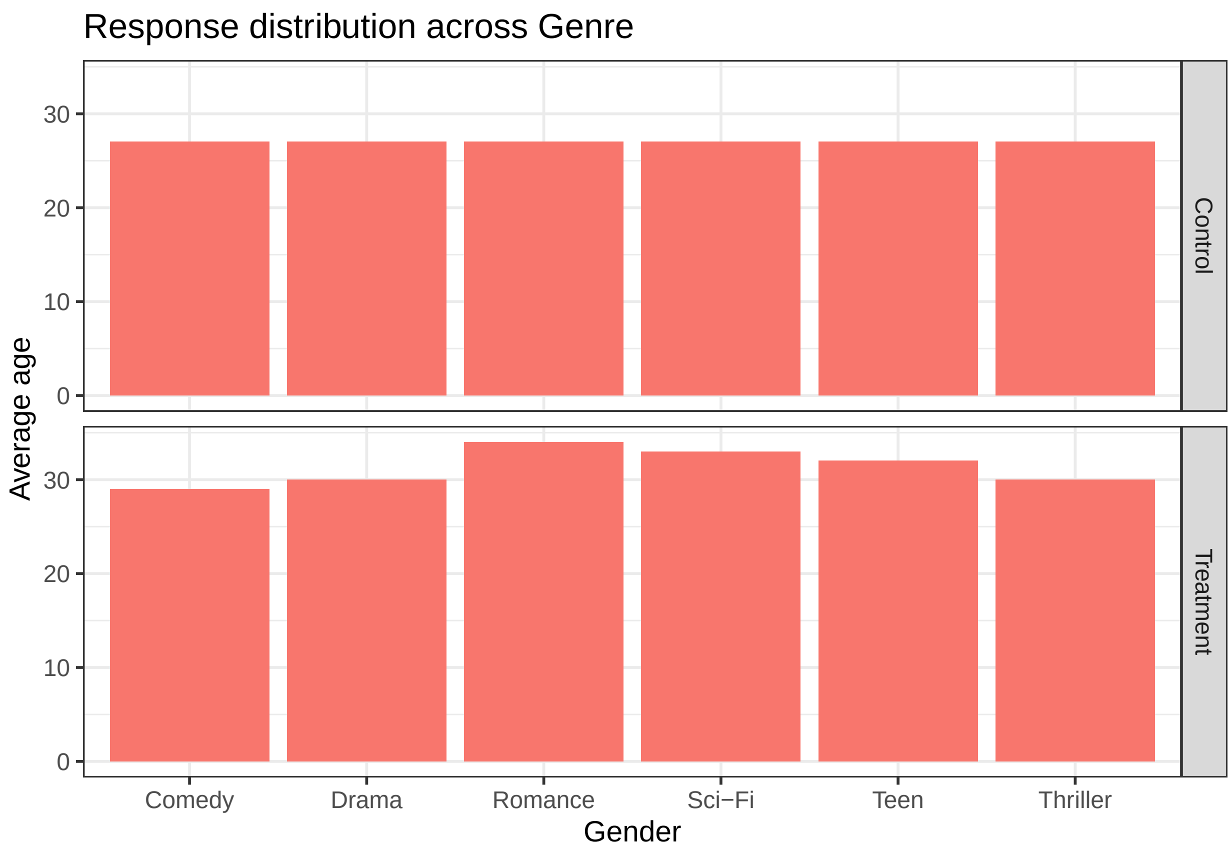
The average age of the respondent was 24.69 years. The average Chinese respondent was 23 years old whereas the average Indian was 26 years old.

```r
data_final %>%
  group_by(Flag, Race) %>%
  summarise(avg_age = mean(Age)) %>%
  ggplot(aes(x = Race, y = avg_age, fill = Race)) +
  geom_bar(stat = 'identity') +
  facet_grid(cols = vars(Flag)) +
  labs(title = 'Age across treatment and control',
       y = 'Average age', x = 'Gender') +
  theme_bw() +
  theme(legend.position = "none") +
  geom_text(aes(label = round(avg_age, 0)), vjust = 2)
```



All 5 genres had a fair equal distribution of responses with 66 responses per genre.

```
ggplot(data_final, aes(x = Genre, fill = '#800d00')) +
  geom_bar() +
  facet_grid(rows = vars(Flag)) +
  labs(title = 'Response distribution across Genre',
       y = 'Average age', x = 'Gender') +
  theme_bw() +
  theme(legend.position = "none")
```



Analysis

We first analyzed the treatment effect on the rankings by regressing the Rankings vs Age and Treatment flag while estimating for the within-effect of each race.

Estimating Treatment Effect on Rank

```
model = glm(up_flag ~ as.factor(Gender) +
              as.factor(Treatment) + as.factor(Race) + Age,
            data = data_final,
            family = 'binomial')
summ(model)
```

```
## MODEL INFO:
## Observations: 350
## Dependent Variable: up_flag
## Type: Generalized linear model
##   Family: binomial
##   Link function: logit
```

```
##
## MODEL FIT:
## <U+03C7>²(4) = 5.68, p = 0.22
## Pseudo-R² (Cragg-Uhler) = 0.02
## Pseudo-R² (McFadden) = 0.01
## AIC = 435.26, BIC = 454.55
##
## Standard errors: MLE
## -------------------------------------------------------
##                               Est.   S.E.   z val.      p
## ------------------------- ------- ------ ------- ------
## (Intercept)                   3.43   2.21     1.55   0.12
## as.factor(Gender)M            0.17   0.26     0.63   0.53
## as.factor(Treatment)1         0.22   0.24     0.93   0.35
## as.factor(Race)Indian         0.28   0.37     0.75   0.45
## Age                          -0.19   0.10    -1.94   0.05
## -------------------------------------------------------
```

We find that the hypothesis test for estimating the coefficient fails significance test and we cannot estimate the true effect of the treatment effect.

We further try to estimate if the two result groups (treatment and Control) are actually statistically different.

t-test and Power Calculation

```
data_test = data_final %>% filter(Flag == 'Treatment')
data_control = data_final %>% filter(Flag == 'Control')
t_test = t.test(data_test$Ranking, data_control$Ranking)
pander(t_test)
```

Table 1: Welch Two Sample t-test: `data_test$Ranking` and `data_control$Ranking`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
1.923	340.8	0.05533	two.sided	3.287	2.994

Here, considering alpha to be 0.1, we observe that we have a significant difference of 0.29 between the two groups. The Test group has a mean Rank of 3.28 vs the Control group which has an mean ranking of 2.99, That means the Test group has actually ranked the images with actors from their own race **Lower** than the images that does not have the characters.

We check if there is enough power to carry out these experiment.

```
power_t_test(n=NULL,type=c("two.sample"), alternative="two.sided",
             delta=0.29,power=0.8,sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##               n = 187.6211
##           delta = 0.29
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The t_test results indicate that the two groups are significantly different and the power test calculations also support the requirement for the number of samples. This means that the two groups actually have different coefficients and there are some unobserved confounds affecting the experiment.

We can infer from the above t-test that the average rank for the treatment is 3.28 and the average rank for the control group for 2.99. The average rank for the treatment group is 5.8% lower than the control group

Limitations

Image Bias

The images provided in the survey had biases such as contrast, setting, color, number of people, etc. We can see examples below:

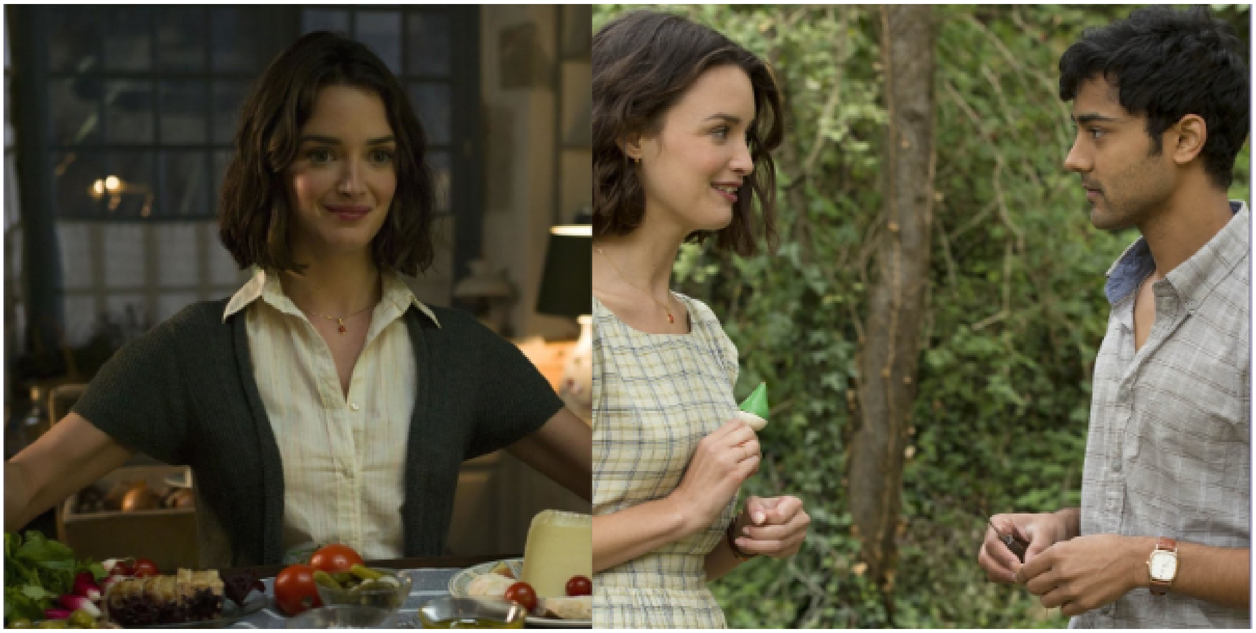


Figure 2: This might be one of the biases which lead to insignificant results observed in the survey these can be controlled by designing more accurate test and control images which are similar to each other

Flawed Survey Design

The Survey contained 6 questions the users had to sit through and judge which can be cumbersome. Each Question had 5 images to rank, which means a total of 30 images to judge, We were restricted by a low number of users because of which we had to send this large volume of images, but this can be easily fixed by making the test pool bigger.

Non-Usable Mobile Interface

The survey was designed on a desktop and worked fine on that platform but the Mobile UI was not optimized for ranking and viewing images. This led to a lot of Mobile surveys being retaken for Desktops or respondents just not putting enough effort to answer the question.

Number of Responses Collected

Even though each respondent had to answer 6 questions, survey design, issues on mobile platforms lead to a lot of responses being invalid. Because of this, we had to carry out a lot of data cleaning and filtering because of which, the dataset size reduced from 480 to 393

Population Demographics

The survey was only circulated within the MSBA cohort, so the analysis results should only be interpreted for demographics similar to the cohort.

Conclusion

From the analysis, we conclude that having a thumbnail from the same race actually has a negative effect on the perception of an unknown show. The two groups were significantly different from each other, but we do not have enough evidence to establish how much is the ranking affected by having a test image in the thumbnail. We conclude this might because of unobserved confounds in the survey relating to pre-existing biases, incorrect design and also the unknown population demographics. Also, the results of this survey are only applicable to the MSBA cohort. The survey can be designed in a better way and experiment can be conducted on a wider audience taking the learnings to get more actionable and concrete results which can aid Netflix to reach a wider adience