Effect of Race on Perceptions of Professionalism

W241 Experiments and Causal Inference

Summer 2020

Ammara Essa | Ryan Sawasaki | Benjamin Silk | Joanna Wang

Abstract

Experiment-based attempts to document racial prejudice are well represented in academic literature. Our study attempts to add to this work by employing a survey-based methodology to test whether race influences perceptions of clothing formality; specifically, we show subjects four images of the same articles of clothing worn by either a Caucasian or Black model, and ask the subjects to rate the clothes' conformity to the standards of "business casual." Our study did not indicate any significant effect of the model's race on perceptions of clothing across the images we tested. In addition, we did not note any statistically significant differences in the perception of subjects of different gender or ethnic groups. As a result, we do not claim to have identified any clear implicit biases in this study.

Background

Recent events have returned race relations to the forefront of American attention. While overt racial discrimination makes national headlines and ignites a wide spectrum of emotions, the majority of people may not experience overt forms of bigotry in their daily lives. More likely to occur are incidents of implicit racial bias, which are subtle, subconscious thoughts and perceptions that are typically shaped through an individual's experiences and social conditioning. Despite good intentions, implicit biases are thought to be pervasive in most people and can be difficult to identify through introspection, and implicit racial bias can lead to prejudiced behavior or, in extreme cases, discriminatory practices and policies at a systemic level.

Research Question

While exploring the broader impact of social conditioning and racial bias in America is a worthwhile study, this experiment was scaled down to fit the time and resource constraints inherent in a class setting. This study examines implicit racial bias within the context of a corporate setting. The research question that this experiment intends to answer is “Does implicit racial bias influence perceptions of other people’s levels of professionalism?” This question will be addressed through an experiment that examines if people’s perception of “business casual” attire changes depending on the skin tone of the person wearing the attire. Formally, we will be testing the null hypothesis of no average effect.

$$\mathbf{H}_0 : \mu_{\mathrm{Y(1)}} = \mu_{\mathrm{Y(0)}}$$

The expectation is that the outcomes will move in a negative direction due to treatment. In other words, it is expected that those with a darker skin tone will be perceived to be dressed less professionally than those with a light skin tone while wearing the same clothing. Although an experiment of this nature will not cure systemic racism within a company, management cannot openly speak on an issue if they are not even aware that one exists. As many companies introspectively re-examine their culture and policies, it may be appropriate to first determine and recognize the extent of implicit racial bias within the company. Indeed, such a result could motivate discussions of implicit racial bias and help the company work towards actionable solutions. It should be noted that we understand there is a distinction between skin tone and racial identity. Skin color is a physical characteristic; race is a complex social construct. However, for the purposes of this experiment, skin tone and race will be used interchangeably, recognizing their commonality but also acknowledging the nuanced differences between the two.

Experiment Design

To test the hypothesis that those with a darker skin tone will be perceived to be dressed less professionally than those with a light skin tone while wearing the same attire, we implement a within subjects randomized experiment.

In this experiment, the treatment is described as an image of an outfit worn by a model with dark skin tone. As such, the control image is the same outfit but with a light skin toned-model. Each participant is randomly assigned to view either the control image or the treatment image. Each image is preceded with the standard industry definition of “business casual” attire and the participant is then asked to rate each image’s conformity to “business casual” standards on one of five levels: strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, or strongly disagree. We specifically chose outfits that may be open to interpretation and avoided anything obviously formal such as a business suit or obviously casual like shorts. We ultimately aim to compare the appropriateness rating of an outfit worn by a light skin toned model to that of the same outfit modelled by a dark skin toned model. The within-subjects aspect comes into play since the same participant is asked four such questions, each for a different attire type. Moreover, the allocation of a participant to the control or experimental group for each attire type is random. We chose a within-subjects design for a few reasons. Most importantly, a within-subjects design allows us to avoid some confounding factors such as each person’s general perception of clothing or some other fundamental differences between participants. Since most participants are exposed to both control and treatment images (albeit in different attire categories) we can capture each person’s ‘baseline’ sentiment and reduce variance between participants when conducting our analysis. Secondly, this type of experiment design helps increase power by obtaining more observations. One participant will respond to four randomly assigned images, thereby giving us four times the number of observations as participants. Lastly, we were also interested to see whether attire types are rated differently based on clothing type and model gender. Figure 1 shows the four categories of attire (control and treatment), and Figure 2 is an example of the survey question setup.

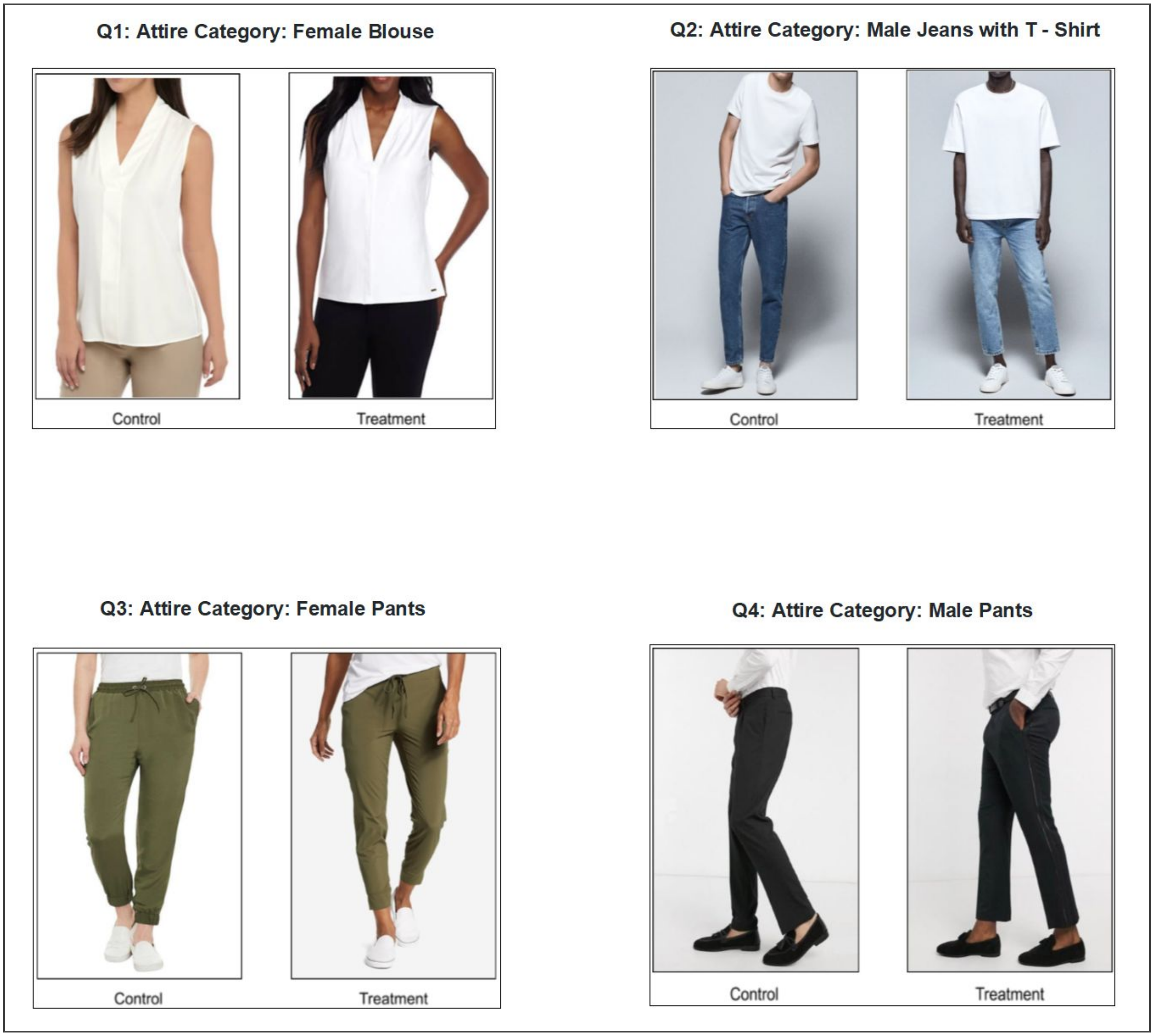
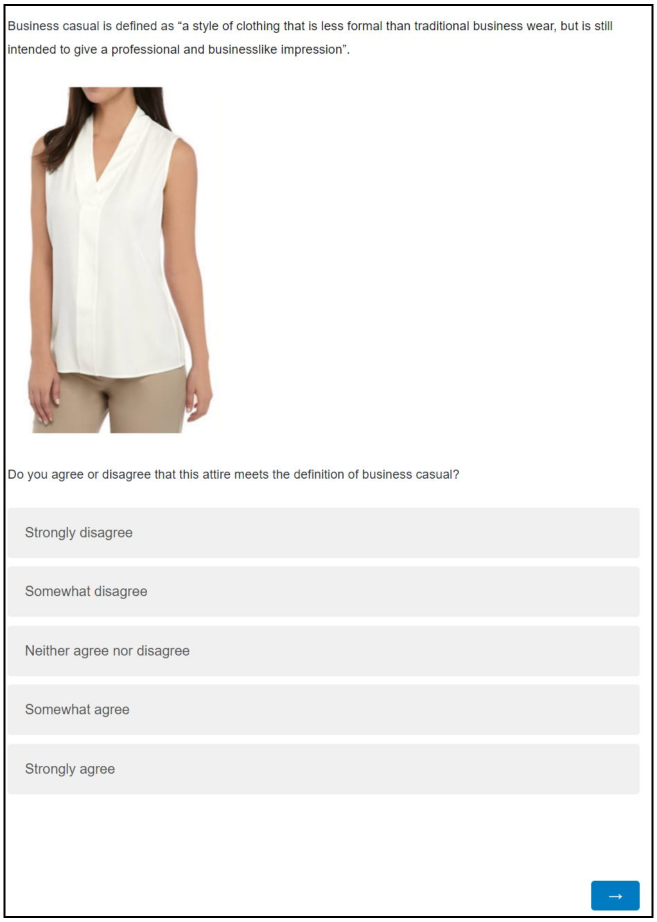

Q1: Attire Category: Female Blouse
Control
Treatment
Q2: Attire Category: Male Jeans with T - Shirt
Control
Treatment
Q3: Attire Category: Female Pants
Control
Treatment
Q4: Attire Category: Male Pants
Control
Treatment

Figure 1 : “Control” and “Treatment” images for each attire type



Business casual is defined as “a style of clothing that is less formal than traditional business wear, but is still intended to give a professional and businesslike impression”.

Do you agree or disagree that this attire meets the definition of business casual?

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

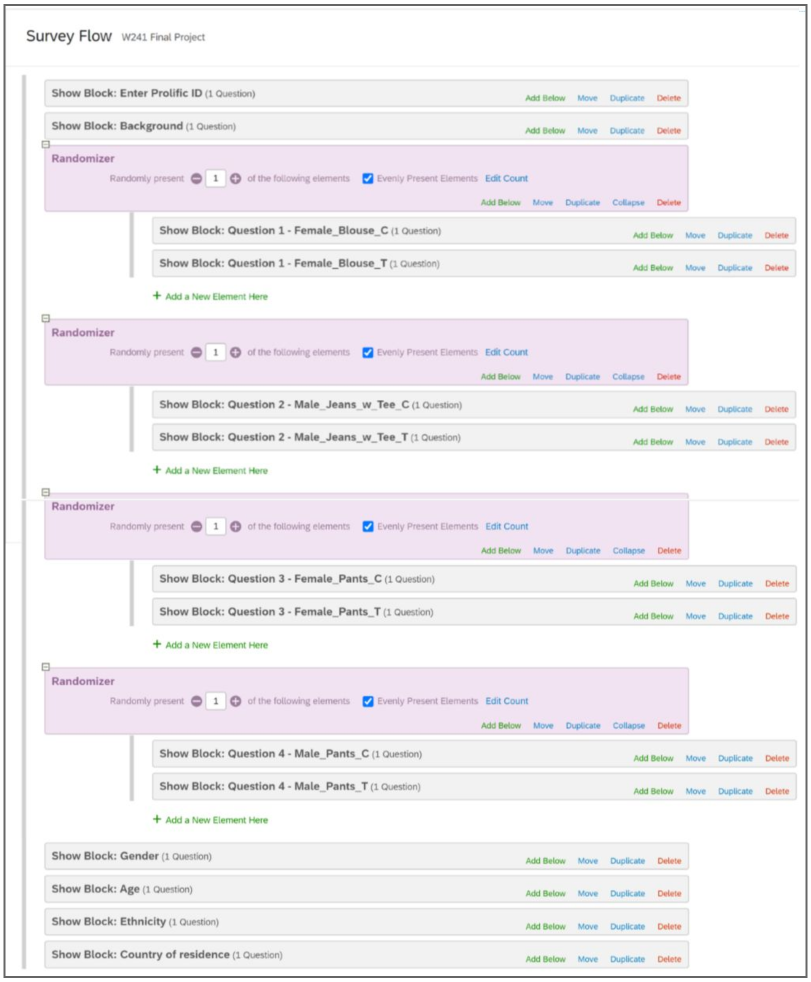
Figure 2 : Example of survey question page setup

Recruitment Process

We distributed our survey via Prolific, a paid online service that recruits participants for studies and surveys. We restricted participation to US residents in order to avoid any additional cultural influences or biases that may skew the results (e.g. workplace attire in foreign countries may be judged differently than in the United States). Since Prolific participants are only paid if they answer every question of the study, we were not very concerned about non-compliance or attrition, and we received complete results for the 460 participants we had requested.

Randomization Process

Qualtrics offers an excellent method of randomization of data displayed to the survey taker. In our implementation, the randomization was done independently for each question such that a participant had an equal chance of being shown the control image as the treatment image for each attire category. The Qualtrics survey flow implementation is also shown in Figure 3. In addition to the survey sentiment response, Qualtrics survey results provide details on exactly which image for each attire category was shown to each participant. During a 1-week pilot study of survey among 30 non-Prolific users, including ourselves, we verified that participants were in fact shown different sequences of attire types every time they took the survey. More specifically, if the same person took the survey again, they would have a 93.75% chance of seeing a different sequence of images.



Survey Flow W241 Final Project

Show Block: Enter Prolific ID (1 Question)

Show Block: Background (1 Question)

Randomizer
Randomly present 1 of the following elements ☑ Evenly Present Elements

Show Block: Question 1 - Female_Blouse_C (1 Question)

Show Block: Question 1 - Female_Blouse_T (1 Question)

Randomizer
Randomly present 1 of the following elements ☑ Evenly Present Elements

Show Block: Question 2 - Male_Jeans_w_Tee_C (1 Question)

Show Block: Question 2 - Male_Jeans_w_Tee_T (1 Question)

Randomizer
Randomly present 1 of the following elements ☑ Evenly Present Elements

Show Block: Question 3 - Female_Pants_C (1 Question)

Show Block: Question 3 - Female_Pants_T (1 Question)

Randomizer
Randomly present 1 of the following elements ☑ Evenly Present Elements

Show Block: Question 4 - Male_Pants_C (1 Question)

Show Block: Question 4 - Male_Pants_T (1 Question)

Show Block: Gender (1 Question)

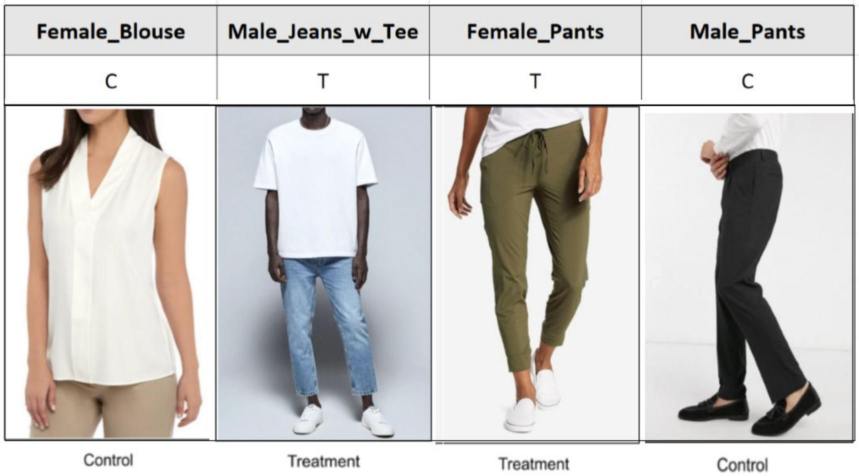
Show Block: Age (1 Question)

Show Block: Ethnicity (1 Question)

Show Block: Country of residence (1 Question)

Figure 3 : Qualtrics survey flow and randomization process

Given our funding limitations, we initiated a respondent request to Prolific for 460 participants. Each participant is randomly assigned the ‘Control’ or ‘Treatment’ version of each attire type. It should be noted that the randomization is independently performed for each picture category i.e. a participant may view the ‘Control’ version (C) of ‘Female Blouse’ and ‘Male Pants’ but receive the ‘Treatment’ version (T) of ‘Male Jeans with T - Shirt’ and ‘Female Pants’. This sequence is illustrated in Figure 4.



Female_Blouse	Male_Jeans_w_Tee	Female_Pants	Male_Pants
C	T	T	C

Figure 4: Example of image sequence displayed to a participant

That said, there are 16 possible sequences in which a survey participant may be displayed the four attire types for feedback as detailed in Table 1. Figure 5 depicts the distribution of image sequences that were observed for the 460 participants and while there are some sequences with higher counts for sequences such as TTCT and TCTC, this imbalance is likely due to the fact that we have a small sample size. Had we increased the number of participants (e.g. N=2000), we would see the image sequence distribution resemble the expected uniform distribution more closely. Moreover, we conducted a chi-square goodness of fit test to compare the observed image sequence distribution to the expected uniform distribution of image sequences. Results from Table 2 show that the p-value = 0.4 of the test is greater than the significance level α = 0.05 and we fail to reject the null hypothesis there is no significant difference between the observed and the expected proportions. Thus we conclude that the randomization was generally successful.

Female_Blouse	Male_Jeans_w_Tee	Female_Pants	Male_Pants	Sequence
C	C	C	C	CCCC
C	C	C	T	CCCT
C	C	T	C	CCTC
C	C	T	T	CCTT
C	T	C	C	CTCC
C	T	C	T	CTCT
C	T	T	C	CTTC
C	T	T	T	CTTT
T	C	C	C	TCCC
T	C	C	T	TCCT
T	C	T	C	TCTC
T	C	T	T	TCTT
T	T	C	C	TTCC
T	T	C	T	TTCT
T	T	T	C	TTTC
T	T	T	T	TTTT

Table 1 : All possible image sequences that can be displayed to participants

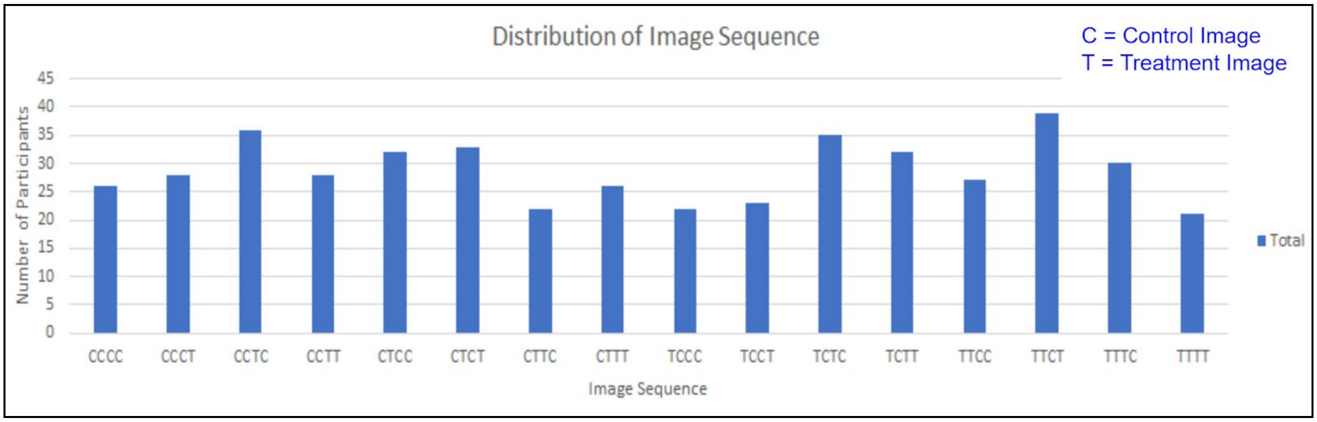
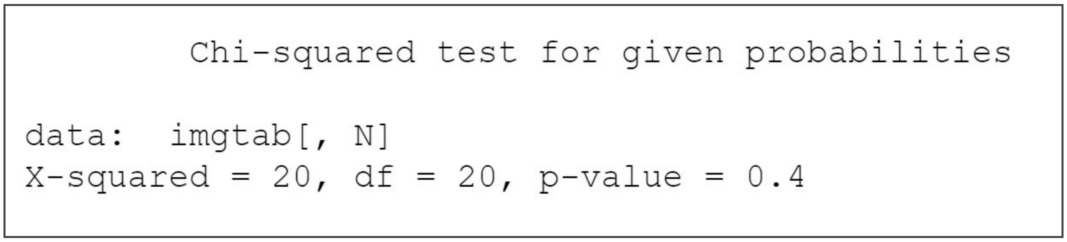


Figure 5 : Distribution of image sequences displayed to survey respondents



```
        Chi-squared test for given probabilities

data:  imgtab[, N]
X-squared = 20, df = 20, p-value = 0.4
```

Table 2 : Chi-square goodness of fit test to compare the observed image sequence distribution to the expected uniform distribution of image sequences

Lastly, the CONSORT flow diagram in Figure 6 documents the sampling, allocation and delivery of treatment amongst participants.

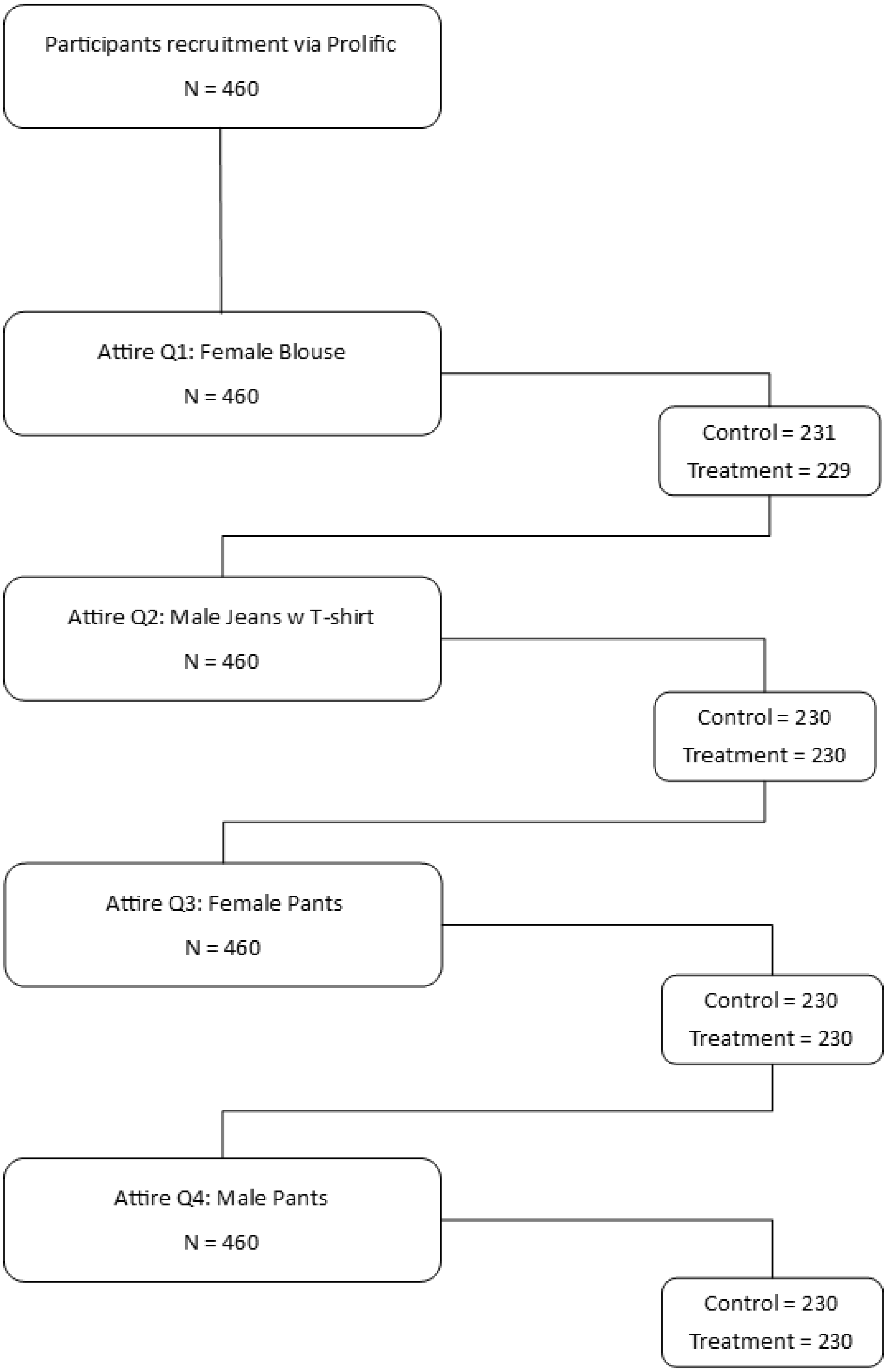


Figure 6 : CONSORT Flow Diagram

Exploratory Data Analysis

After collecting data from Prolific, we looked at the distribution of our participant's gender, ethnicity and age within each attire type as displayed in Figure 7, Figure 8 and Figure 9. We see that the distribution is similar for the control and treatment groups for each attire type. For example, in Figure 7 there are a similar number of male participants and female participants across attire groups. Moreover, within each attire group, there are a similar number of male participants and female participants distributed in the control and treatment groups. Thus we conclude that the random assignment methodology did not result in covariate imbalance in our study.

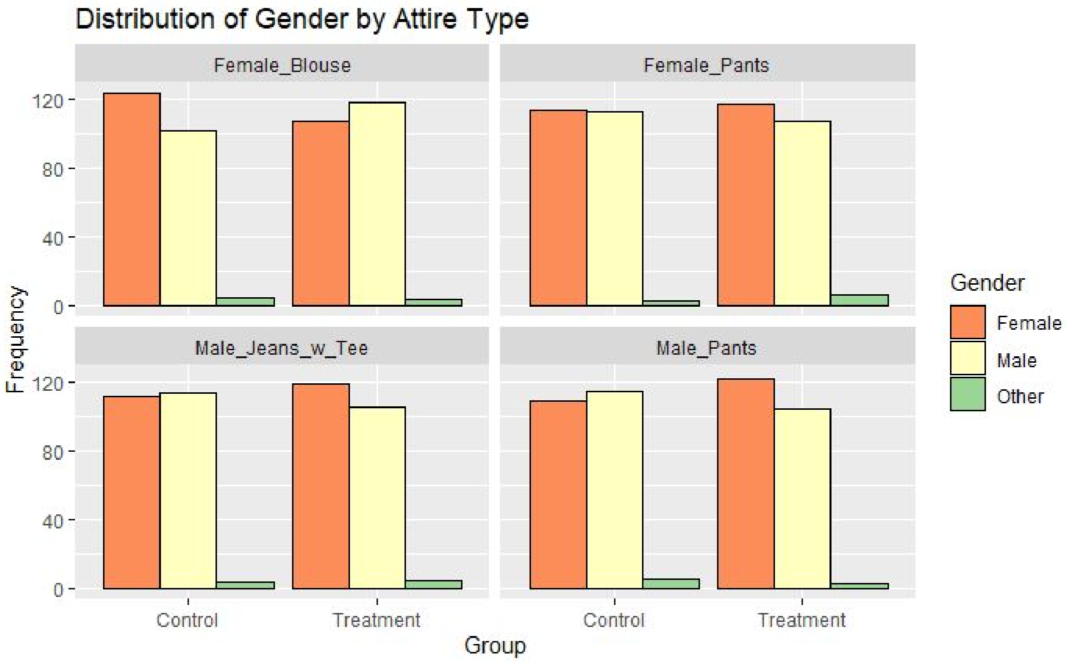


Figure 7 : Distribution of gender by attire type

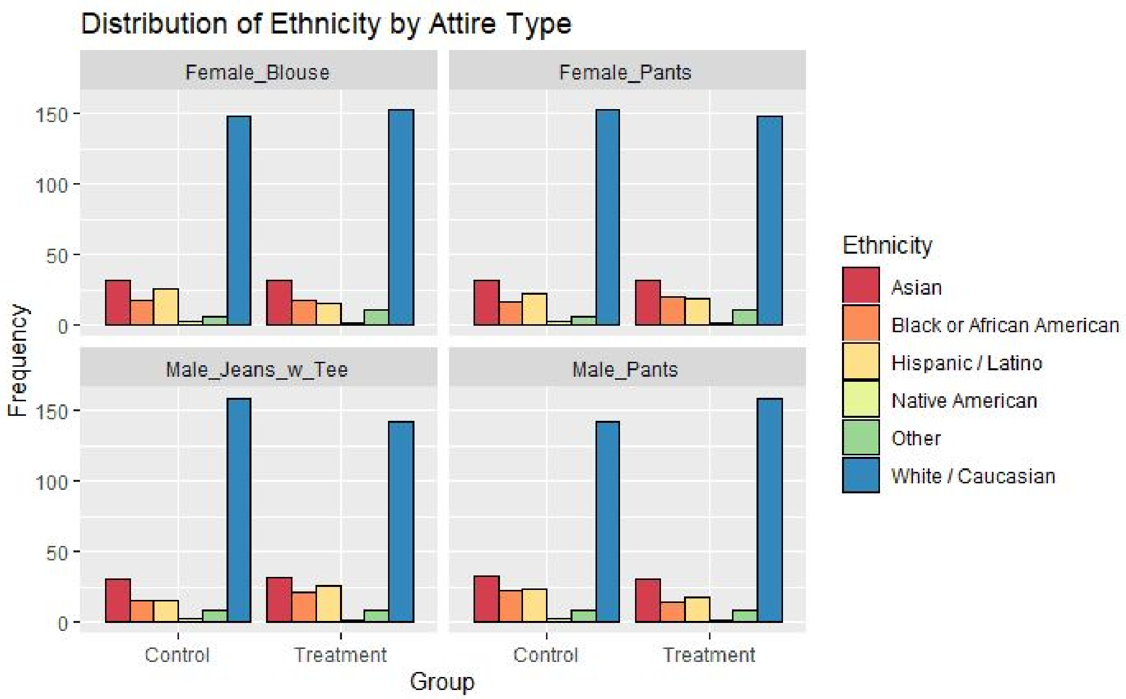


Figure 8 : Distribution of ethnicity by attire type

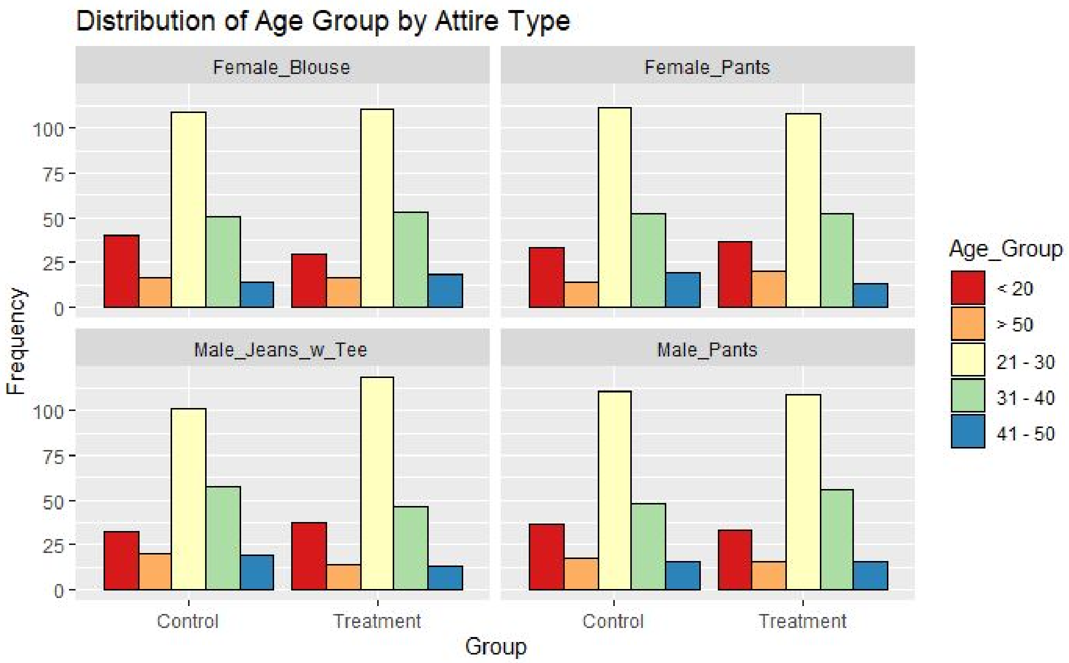


Figure 9 : Distribution of age by attire type

Outcome Measurement

We used Cohen’s d as a measurement of our effect size. Cohen’s d indicates the standardized difference between two distributions that has similar standard deviations and are of the same size. It represents the ratio of the difference between two means divided by the pooled standard deviation. It is mostly used when there are no obvious units that can be used to describe the difference between the two distributions. The formula is:

$$Cohen's\ d = \frac{Mean1 - Mean\ 2}{Pooled\ Standard\ deviation}$$

Cohen’s d fits the purpose of measuring the effect size of our study because: First, the type of measurement we used in our study cannot be easily quantified. In the survey, we asked the participants on a scale of strongly disagree to strongly agree, if they think an outfit is business casual. Then we assigned each statement with a number, which is a proxy but not a precise measurement of the participant’s sentiment. Second, based on our exploratory data analysis results, each attire group has a similar number of treatment and control subjects, which fits the premise of using Cohen’s d. Third, each attire group has similar standard deviation between control group and treatment group, which is the premise for calculation a pooled standard deviation in Cohen's d formula.

To calculate Cohen’s d, we grouped the sentiment result by attire type. Within each attire type group, we grouped the sentiment result by control and treatment. The pooled standard deviation is calculated using the below formula:

$$Pooled\ SD = \sqrt{\frac{Control\ sentiment\ SD^2 + Treatment\ sentiment\ SD^2}{2}}$$

The below table shows our Cohen’s d result for each attire group:

Group	Cohen’s D	Interpretation
Female Blouse	0.063	Trivial Effect
Female Pants	0.413	Small Effect
Male Janes with T-shirt	0.174	Trivial Effect
Male Pants	0.0818	Trivial Effect

Table 3 : Cohen’s d result and interpretation of four attire group’s control and treatment sentiment mean

Based on Cohen's interpretation, we only see a trivial difference between the control group sentiment mean and the treatment group sentiment mean for three attire groups: Female Blouse, Male Janes wither T-shirt and Male Pants, and small difference for one attire group: Female Pants.

Data

Our data consists of the image (treatment or control) that the survey participant viewed for each image pair, the formality level assigned to each image, and the survey taker's gender, age range, and ethnicity. Survey takers rated each image's conformity to "business casual" standards on one of five levels: strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, or strongly disagree.

For the regression analysis, we have chosen to map the 5 level ordinal scale to numerical values ranging from 1 to 5 shown in Table 4. We assume that the difference between any two levels is the same i.e. the difference in sentiment between 'Somewhat agree' and 'Neither agree or disagree' is the same as the difference between 'Somewhat disagree' and 'Strongly disagree'. This allows us to use linear OLS regression to estimate the average treatment effect in numerical terms.

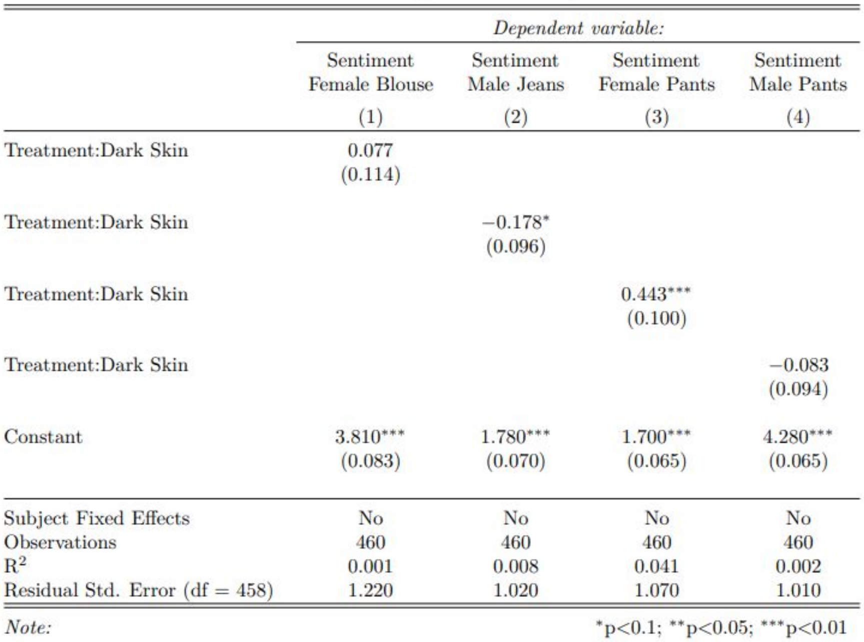
Rating	Numerical mapping
Strongly agree	5
Somewhat agree	4
Neither agree or disagree	3
Somewhat disagree	2
Strongly disagree	1

Table 4 : Mapping ordinal sentiment to numerical values

Models

The first set of models in our analysis estimate the impact of treatment/control group status on clothing ratings for each of the four sets of pictures, as shown in Table 1 below. These models take the form of univariate regressions. The results are varied across the four image sets. For the female blouse and male pants images (denoted by column 1 and 4), we found that subjects' average sentiment for the control group (light-skinned model) is 3.8 and 4.3, respectively, which suggests that the mean respondent somewhat agreed that the clothing meets the definition of business casual when worn by a light-skinned model. In these two images there was a slight positive treatment effect for the female blouse and slight negative effect for the male pants. However, both cases showed no statistical significance. For the male jeans with T-shirt and the female pants images (denoted in column 2 and 3) subjects' average sentiment for the control group is 1.78 and 1.70, suggesting that respondents somewhat disagreed that the attire meets the standard of business casual. The male jeans with T-shirt saw a negative treatment effect at -0.178 at a 10% statistically significant level. The female pants image saw a treatment effect that favored the dark skin model with a treatment effect at 0.443 at a 1% statistically significant level.

Table 5 :OLS Regression of 4 Images

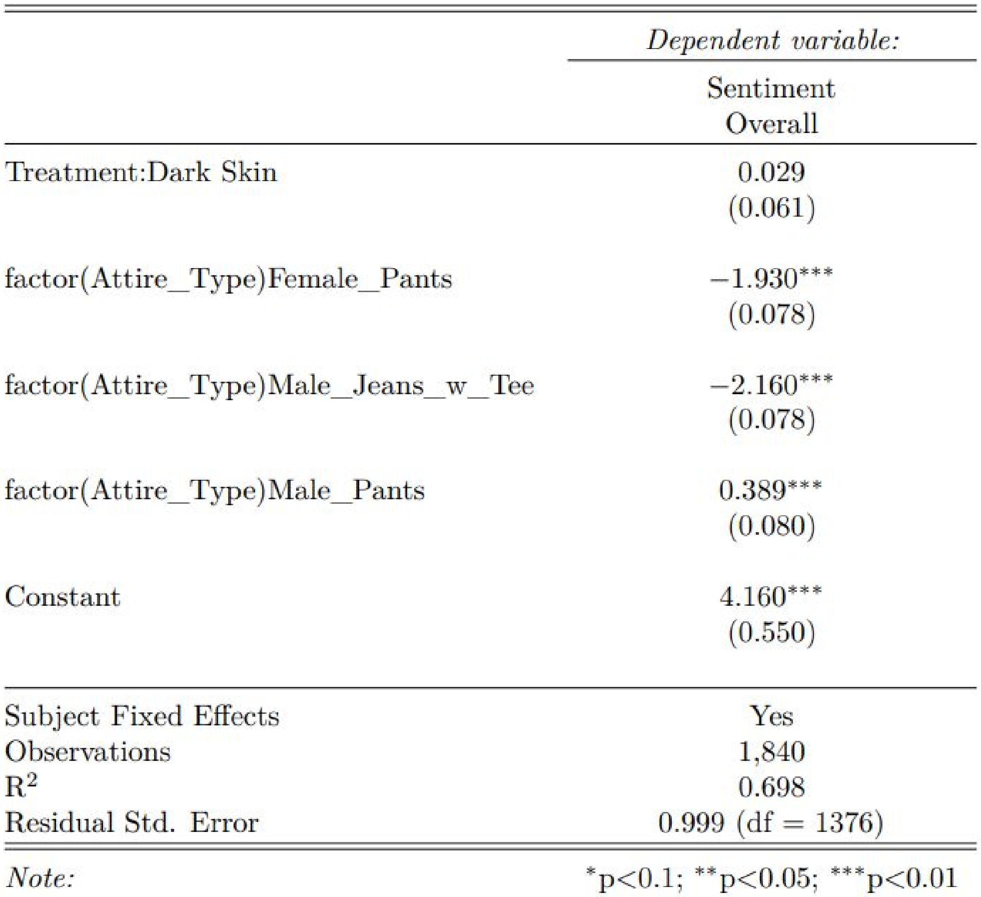


	Dependent variable:			
	Sentiment Female Blouse	Sentiment Male Jeans	Sentiment Female Pants	Sentiment Male Pants
	(1)	(2)	(3)	(4)
Treatment:Dark Skin	0.077 (0.114)			
Treatment:Dark Skin		-0.178^{*} (0.096)		
Treatment:Dark Skin			0.443^{***} (0.100)	
Treatment:Dark Skin				-0.083 (0.094)
Constant	3.810^{***} (0.083)	1.780^{***} (0.070)	1.700^{***} (0.065)	4.280^{***} (0.065)
Subject Fixed Effects	No	No	No	No
Observations	460	460	460	460
R^2	0.001	0.008	0.041	0.002
Residual Std. Error (df = 458)	1.220	1.020	1.070	1.010
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Fixed Effects Testing

The next set of regressions employ subject fixed effects models, which incorporate a fixed effect dummy variable for each of the 460 survey respondents. These models estimate the impact of treatment/control status on sentiment ratings, controlling for the image pair selected and the specific survey-taker conducting the rating. Fixed effects are used because each survey respondent has answered four questions, allowing us to control for respondents' potential tendencies to give predominantly higher or lower ratings across all images. However, these results (Table 2) likewise do not estimate a statistically significant impact of treatment/control status on rating across the entire data set. The use of fixed effects methodology requires us to collapse the four regressions in Table 1 to one regression across the entire data, as the fixed effect measurement requires multiple observations for each survey taker.

Table 6 :OLS Regression Within Subjects

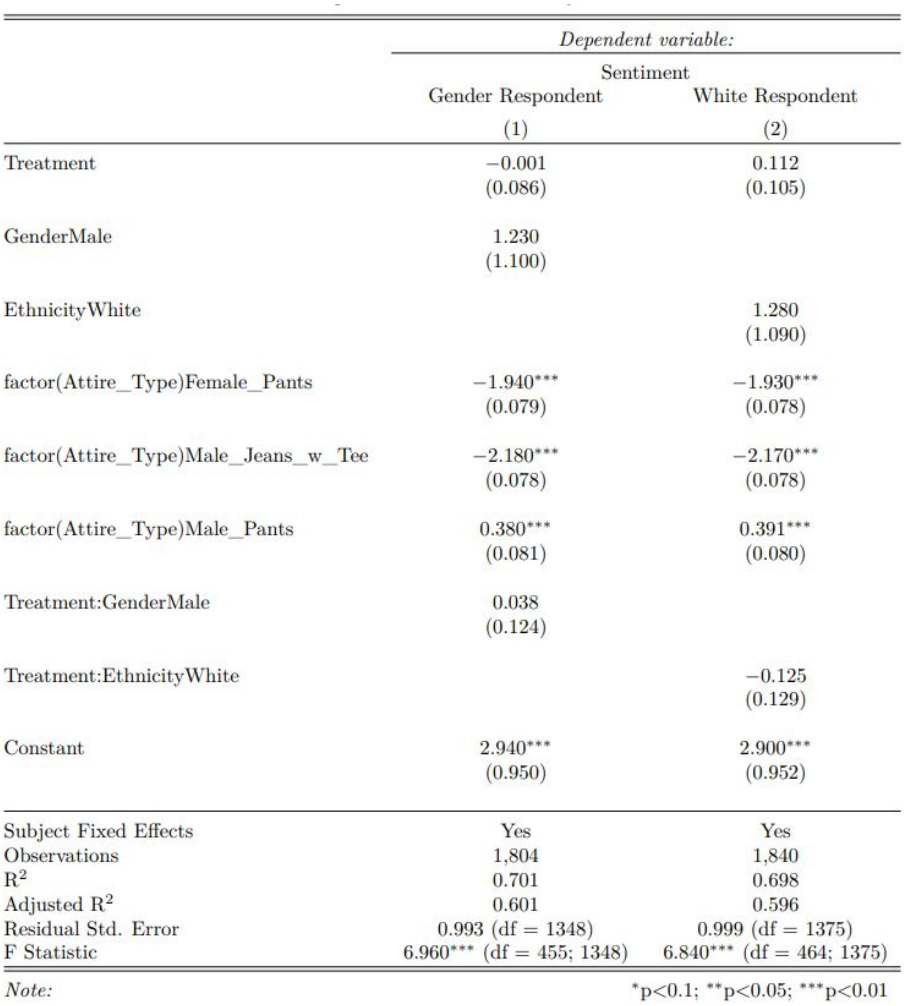


	Dependent variable:
	Sentiment Overall
Treatment:Dark Skin	0.029 (0.061)
factor(Attire_Type)Female_Pants	−1.930*** (0.078)
factor(Attire_Type)Male_Jeans_w_Tee	−2.160*** (0.078)
factor(Attire_Type)Male_Pants	0.389*** (0.080)
Constant	4.160*** (0.550)
Subject Fixed Effects	Yes
Observations	1,840
R^2	0.698
Residual Std. Error	0.999 (df = 1376)
Note:	*p<0.1; **p<0.05; ***p<0.01

Fixed Effects Testing with Heterogeneous Effects

Finally, we tested heterogeneity among respondent groups by estimating regressions with interaction terms between male survey participants and treatment, and white survey participants and treatment to see if the treatment had a different impact on these sub-populations. Maintaining the fixed effects structure, we see that that men don’t find clothing more or less formal than women, that male respondents didn’t have different reactions to Black models than female respondents, that white respondents didn’t judge clothes differently, that white people didn’t react to black models differently at any standard level of statistical significance.

Table 7 :OLS Regression with Within Subject Fixed Effects



	Dependent variable:	
	Sentiment	
	Gender Respondent	White Respondent
	(1)	(2)
Treatment	−0.001 (0.086)	0.112 (0.105)
GenderMale	1.230 (1.100)	
EthnicityWhite		1.280 (1.090)
factor(Attire_Type)Female_Pants	−1.940*** (0.079)	−1.930*** (0.078)
factor(Attire_Type)Male_Jeans_w_Tee	−2.180*** (0.078)	−2.170*** (0.078)
factor(Attire_Type)Male_Pants	0.380*** (0.081)	0.391*** (0.080)
Treatment:GenderMale	0.038 (0.124)	
Treatment:EthnicityWhite		−0.125 (0.129)
Constant	2.940*** (0.950)	2.900*** (0.952)
Subject Fixed Effects	Yes	Yes
Observations	1,804	1,840
R^2	0.701	0.698
Adjusted R^2	0.601	0.596
Residual Std. Error	0.993 (df = 1348)	0.999 (df = 1375)
F Statistic	6.960*** (df = 455; 1348)	6.840*** (df = 464; 1375)
Note:		*p<0.1; **p<0.05; ***p<0.01

Conclusion and Lessons Learned

Our study result failed to reject our null hypothesis that a model’s skin tone does not impact subjects’ perception of the degree to which the model’s clothes conform to the definition of “business casual.” Indeed, we did not find a statistically significant impact of the model’s race on people’s perspective on the outfit formality.

Although a pilot study was conducted to flush out issues and improve upon the experiment, there are a few items that arose during post-experiment that we acknowledge could be improved on for future projects. One of the lessons is the selection of the survey rating system. There were no issues with numerically mapping the sentiments (setting values from 1 to 5 ranging from strongly disagree to strongly agree) and the numerical mapping was useful in simplifying the analysis. However, when presenting and reporting the results, the use of a rating from strongly disagree to strongly agree becomes verbose as we need to describe the original question, the rating system, and the numerical values associated with the rating system. For future experiments, we would consider a survey/rating system that is not only well-suited for analysis, but is also easier to clearly articulate results.

Another lesson that we can carry into future experiments is how to extract even more power from a within subjects design. As noted in our randomization process, each subject received a set of images out of a possibility of 16 sequences. These sequences included the conditions where a subject could receive all treatment or all control images. In conducting the within subjects analysis, more power could have been extracted if we had each subject receive two control and two treatment images, while still randomizing the images.

Finally, we could have phrased our question slightly differently. The research topic of interest was whether models’ skin color impacted the level of their clothes’ perceived formality. However, we phrased the question in terms of degree of conformity to business casual, not realizing that clothes could be either too formal or not formal enough to meet that definition. A question of *“how formal do you consider these clothes to be?”* - scaled from not formal to highly formal - could have addressed our question of interest more directly.