

Pre-Analysis Plan for: Appearance Norms in Academic Settings

Jeffrey Shrader and Tarikua Erda

1. Background

Appearance is potentially important in academic settings: in-person/virtual job-talks for academic hiring are common across many fields; personal websites and videos summarizing research articles are becoming standard; speeches and seminars are a routine part of publicizing papers. Are there definitive norms of appearance in academia, or are they more like unspoken rules and are there differences across disciplines? How does the public's understanding of norms compare against the views of academics? We plan to answer these questions by surveying three groups of respondents—faculty, university students, and lay adults—about their perception of academic job candidates they know to be hypothetical. In particular, we will ask them to rate pictured headshots on perceived competence, professionalism, and agreeability. This study will generate novel evidence on norms of appearance in academic settings.

Among various aspects of appearance, this study will also measure how different hairstyles are perceived and how the perception may vary by race. This is important to study because previous research has shown that there is bias against the natural/Afro-textured hair of Black people, particularly Black women, in labor markets and at the workplace.¹ As such, this research is also relevant to the recent Crown Act (Creating a Respectful and Open World for Natural hair), which recognizes and protects Afro-textured hair/Afro-centric hairstyles (e.g., dreadlocks, also referred to as 'locs') as an immutable aspect of race and/or cultural identity, thereby making hair-based discrimination punishable by law.

2. Sample and Data Sources

Our study aims to examine the perception of appearance. Surveys will be conducted on 3 groups – faculty, university students, and a representative sample of lay adults in the US. Each respondent will rate 32 photos on three different metrics: professionalism, competence, and agreeability. Thus, our total number of observations under each sample will be 32 times the number of respondents. A subset of the photos is not experimentally manipulated and serves only to provide plausible deniability to the respondents. These photos will be excluded from the analysis. On average, 14% of the photos a respondent sees will be these extra photos, and the effective sample size will be the number of experimentally manipulated photos seen by all respondents.

At the time of the filing of this Pre-Analysis Plan, we are planning to launch the lay adult survey (primarily via CloudResearch) and the student survey (primarily via the ORSEE participant recruitment system). We target to accrue approximately 2500 and 500 subjects, respectively. At the time of the filing of this Analysis Plan, only the lay adult and student sample surveys will have been fielded.

¹ See Koval and Rosette (2020) and Opie & Phillips (2015).

We will conduct robustness checks by imposing some sample restrictions. First, we will flag respondents who were below the 3rd percentile or above the 97th percentile in terms of the time it took them to complete the survey and calculate our results with these observations removed. Second, we will conduct robustness checks where we remove the responses of subjects whose response time was less than 1/5 of the median response time. Third, we will also conduct robustness checks where we remove the responses of subjects who failed attention check questions.

3. Outcomes and Relevant Measures

Two primary outcome variables that will be produced from the picture rating activity in our surveys are:

- a) A 10-point Likert scale for models' perceived professionalism, competence, and agreeability. We will use these ratings and analyze, at a high level, how different elements of a photograph (clothing, hairstyle, facial expression, etc) factor into the rating for each of the three metrics.
- b) Additionally, the list experiment in our study will produce a set of primary outcomes:
 - i. The average differences-in-means between the control and treatment groups from the list experiment which saw key sensitive statements about hair: Half of the respondents in our study will be shown 4 statements and asked how many of these they agree with (control group), while the other half will see those same 4 statements plus a key statement about the Crown Act (treated group). We will take the difference in means based on the answers of the control and treated groups using OLS.
 - ii. The share of population that agrees/disagrees with the direct sensitive statements about hair (using the survey data): This is calculated based on the share of population with the sensitive attributes as reflected by responses to a direct question where *all* respondents in the survey are asked to share their opinion on the key statement presented to the treated group in the list experiment.
 - iii. The differences between a) and b): We will take the difference between (a) and (b) to estimate the share of the population which misreports their opinion on the key sensitive statement about hair bias.

Our study will produce the following secondary outcome:

- A quantification of the hair penalty/premium: i.e., the difference across average ratings for photos which depict curly/coily hair or dreadlocks on one hand, and on the other, straight hair (for women) or clean-cut hair (for men). We can calculate the hair premium/penalty as the difference between the average predicted levels of the dependent variable (i.e., the Likert score rating discussed in (3.a) above).

4. Empirical analyses

4.1 Baseline ratings

Using the data from the first half of our surveys (i.e., before any information interventions), we will estimate the below specification. We denote respondent i 's rating of a photo j on the 1–10 Likert scale as V_{ij} . The primary estimating equation will regress V_{ij} on indicators for the race, gender, and hair style of the models, as well as the interactions of these variables. This will involve, for each of the three metrics, variations of the following regression specification.

$$V_{ij} = \beta_1 B_j + \beta_2 G_j + \beta_3 H_{2,j} + \beta_4 H_{3,j} + \beta_5 B_j G_j + \beta_6 B_j H_{1,j} + \beta_7 B_j H_{2,j} \quad (1) \\ + \beta_8 B_j H_{3,j} + \beta_9 G_j H_{1,j} + \beta_{10} G_j H_{2,j} + \beta_{11} G_j H_{3,j} + \beta_{12} B_j G_j H_{1,j} \\ + \beta_{13} B_j G_j H_{2,j} + \beta_{14} B_j G_j H_{3,j} + x'_{ij} \theta + \alpha_i + \varepsilon_{ij}$$

The main right-hand-side variables are an indicator for whether the photo depicts a Black model (B), an indicator for whether the model is female (G), three indicators for the different hair styles that the models can wear (H_k for $k = 1, 2, 3$ for three types of hair styles, describe below). The main coefficients of interest are the third-level interactions which will allow for estimates of the hair penalty for Black women from natural curly hair/afro (H_1) and locs (H_3) versus straight hair styles. Therefore, key hypothesis tests will be whether the coefficient $\beta_{12} < 0$ or $\beta_{14} < 0$. These are two separate tests to determine whether there is a penalty associated with natural afro hair or locs for Black women. Given that we will conclude that there is evidence of a penalty for not wearing straight hair if either coefficient is negative, we will also estimate this specification with a control for the familywise error rate using a stepdown procedure.

The other variables in the regression are controls. These include a subject-level fixed effects α_i and, in additional specifications, controls for other photo attributes (represented by x): an indicator for whether the model is wearing a suit jacket (C), and indicator for whether the photo is in high resolution (R), and an indicator for whether the model is smiling (E). We are also interested in studying other control variables such as the duration the spent spent on each photo and the number of photos the subject had rated before the current photo. In the main specifications, we will estimate the equation without these controls (aside from subject fixed effects) and with the controls as selected through machine learning (see the section on algorithmic control selection below).

We will also estimate a version of the equation that involves levels and interactions for all photo characteristics (race, gender, hair, clothing, expression, and resolution). In additional analyses, we plan to conduct heterogeneity analyses by subject race, gender, political leaning, and views about higher education system.

Using the coefficients from the interaction terms in the above regressions, we will calculate, for each race-gender combination of the pictured models, the average predicted rating for each combination of characteristics (such as clothing, hairstyle, and facial expression) as well as the average standard error associated with each averaged prediction under each metric on which respondents rated the photographs (i.e., for professionalism, competence, and agreeability). This exercise will allow us to understand what elements are most associated with increasing perception of professional/competence/agreeability.

In particular, to understand how hairstyles may factor into perceptions, we will compare the difference between the average predicted rating for each model when pictured in curly hair or

dreadlocks vs in more styled hair (i.e., straight hair for women and clean-cut hair for men), and again compare the sizes of the difference across models of the same gender but different races. This enables us to develop a key measure: the hair penalty/premium as the difference across average ratings for photos which depict a less styled natural curly hair or in locs as compared to more styled hair. E.g., if the average rating for photos depicting more styled hair is higher than that for photos depicting less styled curly hair for the same model, then we will conclude that there is a penalty for curly hair/less styled hair for populations of that model's race and gender; if it is lower, then the implication would be that there is a premium on low-effort hair for populations of that model's race and gender.

4.2 Treatment effects

Our study involves two different information treatments that will be introduced at the half point of the survey to randomly selected thirds of the subjects (randomized at the subject level). The first treatment is a vignette that tells respondents that authenticity at work boosts productivity (Generic Treatment) while the second one is a message about the legacy of hair discrimination in the US (Historical Explanation Treatment). Additionally, to address experimenter demand on the two groups of respondents who received either one of the two information intervention treatments, we will follow de Quidt et al (2018) and conduct an experimenter demand induction treatment whereby treated respondents will receive no demand induction (with 50% chance), weak demand induction (with 25% chance), and strong demand induction (with 25% chance).

To understand if and how our information intervention treatments affected respondents' ratings, we will re-estimate Equation (1) again for each of the two information intervention treatments. In particular, for each treatment, we will estimate the difference-in-differences in ratings between the control group and the treatment group before and after treatment. This will be done by interacting an indicator variable for each treatment group and the post treatment indicator with the attributes from Equation (1). For instance, Equation (2) below interacts both the treatment status $Treat^{Hist}$ (which takes the value of 1 if respondent i received the Historical Explanation Treatment before the 2nd round of ratings and 0 if she was in the control group) and a $Post$ dummy (for the second half of ratings) with each term. The underlying sample for the specification in Equation (2) would be the observations from the control group and the Historical Explanation treatment group. Similarly, we would use a sample restricted to only the control group and the Generic Treatment group to separately estimate an equation in which we replace $Treat^{Hist}$ by $Treat^{Generic}$, which would take the value of 1 if the respondent received the Generic vignette treatment and 0 if she was in the control group.

$$\begin{aligned}
 V_{ij} = & \beta_1 B_j * Treat^{Hist} * Post + \beta_2 G_j * Treat^{Hist} * Post + \beta_3 H_{2,j} * Treat^{Hist} * Post + \\
 & \beta_4 H_{3,j} * Treat^{Hist} * Post + \beta_5 B_j G_j * Treat^{Hist} * Post + \beta_6 B_j H_{1,j} * Treat^{Hist} * Post + \\
 & \beta_7 B_j H_{2,j} * Treat^{Hist} * Post + \beta_8 B_j H_{3,j} * Treat^{Hist} * Post + \beta_9 G_j H_{1,j} * Treat^{Hist} * Post + \\
 & \beta_{10} G_j H_{2,j} * Treat^{Hist} * Post + \beta_{11} G_j H_{3,j} * Treat^{Hist} * Post + \beta_{12} B_j G_j H_{1,j} * Treat^{Hist} * \\
 & Post + \beta_{13} B_j G_j H_{2,j} * Treat^{Hist} * Post + \beta_{14} B_j G_j H_{3,j} * Treat^{Hist} * Post + x'_{ij} \theta * Treat^{Hist} * \\
 & Post + \alpha_i + \varepsilon_{ij}
 \end{aligned} \tag{2}$$

The estimates from this exercise will enable us to study how the average predicted ratings for each race-gender-hairstyle may change, if at all, after each of the two treatments.

To maximize power, we will first estimate the above equation by pooling together all respondents who received each information treatment. Additionally, to be able to bound potential effects of experimenter demand, we will estimate Equation (2) separately for each demand induction level (no, weak, or strong) using the respondents who did not receive any information intervention treatments as a control group. In additional analyses, we also plan to analyze treatment effect heterogeneity by respondent characteristics, including race and gender, as well as political leaning, and views about higher education system.

4.3 Algorithmic control selection

We will follow Ludwig, Mullainathan, and Spiess (2019) in implementing a machine learning-based control procedure when estimating out treatment effects and preferences. In particular, we plan to use LASSO to algorithmically choose which controls to include for power maximization.

4.4 List experiment

To identify and quantify potential social desirability bias associated with expressing negative views about curly hair or dreadlocks about Black individuals, we will use the list experiment questions in our survey and follow the approach of Aksoy et al (2022). First, using respondents' answers to the list experiments, we will use OLS to identify the difference-in-means between the 'control' group (the group that only saw 4 non-sensitive statements) and the 'treated group' (the group which saw the same 4 statements plus an additional key statement about the Crown Act). This difference will be captured by β_1 , which is shown in equation (2) below. Then, we will also calculate the share of population that agree with the same key statement that is presented later in the survey when it is posed as a direct (non-veiled) question (%share_direct_question). Lastly, we will take the difference between β_1 and the population share from the direct question: if the share of population identified by the direct question is lower than that identified by comparing the veiled responses, then the difference between those two quantities (i.e., $\beta_1 - \text{\%share_direct_question}$) is the share of the population that only feels comfortable with revealing agreement with the 'key' statement anonymously but not directly. In other words, this will indicate to us $(\beta_1 - \text{\%share_direct_question}) * 100$ percent of the population exhibits social desirability bias and thus underreported agreement with the key statement in the direct question.

$$\#_of_agreed_with_statements_i = \beta_0 + \beta_1 \text{Saw_key_statement} + \beta_2 X_i + \varepsilon_{ij} \quad (3)$$

Null hypothesis to test: $\beta_1 = 0$

We will estimate this specification in the above naïve form as well as by including control variables, X_i , for respondent baseline (pre-treatment) characteristics using the machine learning approach outlined in Section 4.3.

5. Survey attrition

We will report statistics on how many respondents did not complete the survey and how far into the survey they got on average before they left the survey. If any respondents got far enough into the survey to have filled out demographic information, we will report summary statistics and study whether there is any statistically significant associated between attrition and demographic characteristics using an omnibus F-test. Currently, we do not have any prior ex-ante why attrition in this survey may be systematically driven by any characteristics of respondents in a way that may affect our findings.

6. References

Aksoy, Billur, Christopher S. Carpenter, and Dario Sansone. *Understanding Labor Market Discrimination Against Transgender People: Evidence from a Double List Experiment and a Survey*. No. w30483. National Bureau of Economic Research, 2022.

de Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and Bounding Experimenter Demand. *American Economic Review*, 108(11), 3266–3302. <https://doi.org/10.1257/aer.20171330>

Koval, C. Z., & Rosette, A. S. (2020). The Natural Hair Bias in Job Recruitment. *Social Psychological and Personality Science*. <https://doi.org/10.1177/1948550620937937>

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess. "Augmenting pre-analysis plans with machine learning." *AEA papers and proceedings*. Vol. 109, 2019.

Opie, T. R., & Phillips, K. W. (2015). Hair penalties: The negative influence of Afrocentric hair on ratings of Black women's dominance and professionalism. *Frontiers in Psychology*, 6, Article 1311. <https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01311/full>