Supplementary Appendix for "Realistic Image Primes for Experimental Research"

Justin de Benedictis-Kessner* Tess Wise†

August 23, 2021

Contents

*Assistant Professor, John F. Kennedy School of Government, Harvard University, jdbk@hks.harvard.edu. Twitter: @jdbk.

†Assistant Professor, Department of Politics and International Affairs, Wake Forest University, wiset@wfu.edu. Twitter: @tesswise.

A Image Collection Process

In this section, we describe the process by which we gathered images for our database and standardized them for use in experiments.

The images in RIPER were all drawn from the websites of large employers, including universities, that publicly posted photos of their employees and students, as well as political candidate webpages from local electoral contests between 2012 and 2015. When collecting images from a certain community (such as one department at one university) we gathered all images on that department's site – rather than rebalancing the images we collected to better represent the academic discipline or the United States or the global community along gender or racial lines. Our image database was not constructed to be exhaustive (i.e. all people who are academics, or all politicians) but to provide a broad variety of images of people that are members of certain professional communities. The totality of the images included in our database are not meant to be a representative picture of any of the professional groups or racial groups on which we targeted our data collection efforts. As researchers, we have a responsibility not to reinforce stereotypes about race, ethnicity, and gender, but do acknowledge that the databases provide a picture of these disciplines that may perpetuate stereotypes because of their lack of diversity within professional groups. However, our goal is to enable the scholarly study of dynamics of race and gender, as well as many other theories tied to these demographic concepts. To enable this type of inquiry, our discipline needs more tools to study how racial and gender stereotypes structure people's attitudes and behavior. As such, we provide this database to the research community as a convenience sample of images without any claim that it is representative along race or gender lines.

None of the photos in our database were pulled from social media accounts or private websites, and none were posted with explicit copyright statements. Our understanding, after conferring with our universities' legal offices and research librarians, is that our re-use of photos posted on public websites as headshots (which is how all of our images were posted) for research purposes is governed by fair use doctrine. Fair use rights, while subject to some restrictions, are a flexible concept that allows for the use of material – even copyrighted material – under some circumstances, including classroom teaching or scholarly research. The factors considered in determining fair use rights are:[1]

1. **The purpose and character of re-use.** Given that our purpose for re-using the images in the RIPER database is for academic research in which the images convey information (about various demographic characteristics) and is accompanied by significant original content other than the images (i.e. the paper's text, along with the ratings of the images), we believe that our use of these images falls under an acceptable purpose for fair use to apply.

2. **The nature of the copyrighted work.** None of the images which we compiled into the RIPER database were posted on websites with explicit copyright statements. Nor were any of these images posted with a nature that is creative, limiting the application of copyright to them.

[1] `https://www.copyright.gov/fair-use/more-info.html`

3. **The amount and substantiality of the portion of content taken.** Since we use a mixture of either the entirety of the images without alteration (i.e. not cropping them) or some portion of the images (cropped images) that we gathered from websites, it can be considered fair use to use them in an educational or research setting.

4. **The effect of the use upon the potential market.** None of the images in RIPER were posted on commercial websites with any discernible market value for the images. Furthermore, use of these images in the RIPER database has no commercial impact given that we are not using the database for any profit-driven purpose.

We believe part of the contribution of RIPER is its value for use by a broad number of academic researchers. We do not intend for the use of the images in RIPER to be re-used for any profit-driven purpose. As explained by the Visual Resources Association in their summary of fair use of images in educational settings (including research), the use of images and their distribution to others in the scholarly community is best positioned to fall under fair use doctrine if the people distributing the images make good faith efforts to notify users of the images (i.e., other researchers) that the images are being made available for teaching, study, or research only.[2] This motivated us to make use of our image database contingent upon an agreement by those using it to only do so for teaching, study, or research.

[2] `https://vraweb.org/wp-content/uploads/2011/01/VRA_FairUse_Statement_Pages_Links.pdf`

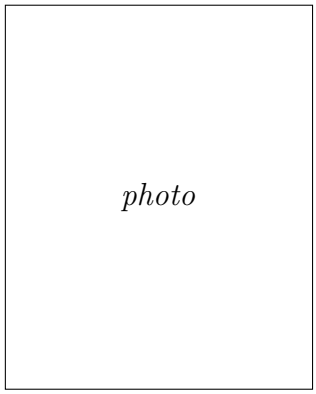
B Rating Task Questions

Introduction text:
We are interested in your first impressions of people. On the coming pages you will see a series of faces. We will ask you to judge:

- how attractive they appear to you
- what race or ethnicity you believe they are
- how much money you think that they make in a year

There are no right or wrong answers. It's important that you simply provide the answer that is closest to your "gut feeling."

Attractiveness evaluation:
How attractive do you think the following person is, if 0 is not attractive at all and 10 is highly attractive? [Slider: 0 to 10]



Perceived income:
What do you think the above person's household's annual income is?

- Less than $10,000
- $10,000-$14,999
- $15,000-$19,999
- $20,000-$24,999
- $25,000-$29,999
- $30,000-$39,999
- $40,000-$49,999
- $50,000-$59,999

- $60,000-$69,999
- $70,000-$79,999
- $80,000-$99,999
- $100,000-$119,999
- $120,000-$149,999
- $150,000 or more
- Prefer not to say

Perceived race:
What race do you think the above person is?

- White/Caucasian
- Hispanic or Latino
- Black/African-American
- Asian
- Other _____

C Rater Population Characteristics

In this section, we describe the demographic characteristics of the MTurk workers that we employed to report their perceptions of the images in our database. As shown in Table A1, the raters were approximately 40% female, and while not representative of the U.S. (or world) population along racial lines, from a variety of racial/ethnic backgrounds, were a diverse set of ages, and represented all geographic regions of the U.S.

Despite the fact that our raters do not necessarily match the demographics of the US population, the danger that might invalidate the use of the images in our database is, rather, that our ratings of the images in the RIPER database do not necessarily match the ratings they might receive from a representative sample of the US or global population. However, previous research examining experiments conducted on MTurk and comparing them to experiments conducted on other sample pools would suggest that this danger is unlikely (e.g. Berinsky, Huber, and Lenz, 2012; Coppock, 2019). In line with these studies, we would expect that the MTurk population's ratings are more likely to match ratings from a representative group of US residents than other populations typically used to pretest images, such as researchers themselves, research assistants, or university student subject pools.

Table A1: Demographics of MTurk workers

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Female	1,447	0.396	0.489	0.000	0.000	1.000	1.000
White	1,447	0.759	0.428	0.000	1.000	1.000	1.000
Black/African-American	1,447	0.048	0.215	0.000	0.000	0.000	1.000
Hispanic/Latino	1,447	0.059	0.237	0.000	0.000	0.000	1.000
Asian	1,447	0.105	0.307	0.000	0.000	0.000	1.000
Age	1,447	32.428	10.186	18.000	25.000	37.000	78.000
Region: Midwest	1,442	0.229	0.420	0.000	0.000	0.000	1.000
Region: Northeast	1,442	0.238	0.426	0.000	0.000	0.000	1.000
Region: South	1,442	0.297	0.457	0.000	0.000	1.000	1.000
Region: West	1,442	0.236	0.425	0.000	0.000	0.000	1.000

D Database Features

The RIPER Database contains 249 photographs. Each photograph has the following 44 features associated with it. Data on these characteristics can be found in the `ratings_summary.csv` file of the replication data archive.[3]

While the images in RIPER are meant to provide data on several perceived characteristics of the people depicted in the images (attractiveness, income, and race), along with several machine-rated characteristics using Microsoft Azure Cognitive Services API (estimated age, gender, head pose in three dimensions, smile, makeup, glasses, hair color, baldness, eight indicators of emotion, and three indicators of facial hair), and researcher-rated characteristics (jewelry, necktie, a 0/1 indicator of facial hair, a 3-level indicator of expression, neutral background, studio background, background color and objects, professional headshot, and a text field for other features such as the presence of lanyards or scarves). The ratings presented in the summary file are by no means meant to be exhaustive descriptions of each image. There are an almost infinite number of characteristics which might distinguish the photos of people in our database, and the ratings we present are only one set of characteristics that might matter when using the images in RIPER to test theories of social science.

As a result, we suggest that social scientists using images in RIPER take these ratings as a starting point to identify potentially similar photos to use in experiments. If there are strong theoretical reasons to believe that another characteristic not included in the summary ratings of images in RIPER, such as femininity or masculinity of the image, matters for the research outcome of interest, we suggest additional pre-testing of a subsample of images in RIPER along those dimensions before using them. As with any part of the experimental research process, we discourage careless use of any image in RIPER without due attention to the numerous theoretical constructs that it may vary alongside the construct of interest.

Features of images (variable name):

1. **Name** (`name`)
   The four or five character code that uniquely identifies the photograph. The first character is an indicator of professional group (a = Academic, g = Graduate Student, p = Professional, w = Working Class). The second character is an indicator of the racial group (b = Black, w = White, l = Latino). The third character is an indicator of gender (m = male, f = female). The final character(s) is a random number that allows the photograph to be uniquely identified.

2. **Professional Group** (`prof_group`)
   The professional group that the photograph comes from. Professional may be `academic`, `grad student`, `professional` (politicians), or `working class` (facilities workers).

3. **Racial or Ethnic Group** (`race_group`)
   The racial or ethnic group assigned by the research team. Racial or ethnic group may be `black`, `white`, or `latino`.

[3]One photo, "gbf8," was not machine-readable by artificial intelligence used to calculate some measures and consequently does not have values for the machine-generated features.

4. **Gender** (`gender`)
   The gender of the person in the photo. Gender may be male (`m`) or female (`f`).

5. **Mean Attractiveness** (`mean_attr`)
   The average attractiveness over all ratings, collected through Amazon Mechanical Turk, on a scale of 0 (not attractive at all) to 10 (highly attractive)

6. **Standard Deviation of Attractiveness** (`sd_attr`)
   The standard deviation of attractiveness ratings.

7. **Proportion White** (`prop_white`)
   The proportion of ratings, collected through Amazon Mechanical Turk, that described the person in the photo as white.

8. **Proportion Latino** (`prop_latino`)
   The proportion of ratings, collected through Amazon Mechanical Turk, that described the person in the photo as Latino.

9. **Proportion Black** (`prop_black`)
   The proportion of ratings, collected through Amazon Mechanical Turk, that described the person in the photo as black.

10. **Proportion Asian** (`prop_asian`)
    The proportion of ratings, collected through Amazon Mechanical Turk, that described the person in the photo as Asian.

11. **Proportion Other** (`prop_other`)
    The proportion of ratings, collected through Amazon Mechanical Turk, that described the person in the photo as some other race.

12. **Proportion Low Income** (`prop_inc_lt30k`)
    The proportion of ratings, collected through Amazon Mechanical Turk, that perceived the person in the photo as having an income less than $30,000 per year.

13. **Proportion Middle Income** (`prop_inc_30k79k`)
    The proportion of ratings, collected through Amazon Mechanical Turk, that perceived the person in the photo as having an income between $30,000 and $79,999 per year.

14. **Proportion High Income** (`prop_inc_gt80k`)
    The proportion of ratings, collected through Amazon Mechanical Turk, that perceived the person in the photo as having an income greater than $80,000 per year.

15. **Number of Ratings** (`n_rating`)
    The number of individual raters who rated the photo on Amazon Mechanical Turk.

16. **Facial expression** (`expression`)
    A three-level indicator of facial expression (smiling, neutral, or frowning) as coded by the researchers.

17. **Facial hair** (`facial_hair`)
    The presence of facial hair (values: 1/0) as coded by the researchers. This indicator is intended to be less sensitive than the machine-perceived indicators of facial hair (see below) and indicate significant facial hair (beyond, say stubble or "five o'clock shadow").

18. **Presence of neck tie** (`necktie`)
    Presence of neck tie (values: 1/0) on the person in image as coded by the researchers.

19. **Presence of jewelry** (`jewelry`)
    Presence of neck or ear jewelry (values: 1/0) on the person in image as coded by the researchers.

20. **Neutral background** (`neutral_background`)
    An indicator for a professional, studio, or otherwise neutral non-variable background in the photo such as blurry greenery as coded by the researchers (values: 1 if the photo has a neutral background; 0 if the background is more variable).

21. **Studio background** (`studio_background`)
    An indicator for whether the person is posed in front of a studio background as coded by the researchers (values: 1 if posed in front of studio background; 0 if posed in front of a more varied background; -1 if it is unclear).

22. **Professional headshot** (`professional_headshot`)
    An indicator that captures whether a photo appears to be professionally taken as opposed to a vacation photo, or cropped from a casual photo with friends as coded by the researchers (values: 1 if professional; 0 if not; -1 if unclear).

23. **Background color and objects** (`background_description`)
    An text description of the background of the photo and any discernible objects as coded by the researchers (e.g., "yellow, books").

24. **Other image info** (`other_info`)
    Anything else the database creators noticed as a prominent or distinguishing feature of the image such as a lanyard, bow tie, tinted glasses, etc.

25. **Machine-estimated age** (`machine_age`)
    The age in years of the person in the photo, as estimated by Microsoft Azure Cognitive Services' computer vision API.[4]

26. **Machine-calculated head pose** (`machine_head_pose`)
    The position of the head in three dimensions as calculated by Microsoft Azure Cognitive Services' computer vision API. The three numbers are roll, yaw, and pitch angles in

[4]For more details on this service and the algorithm behind it, see `https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-detecting-faces`.

degrees, which are defined according to the right-hand rule. The order of three angles is roll-yaw-pitch, and each angle's value range is from -180 degrees to 180 degrees.[5]

27. **Machine-perceived smile** (`machine_smile`)
    The smile expression of the face between 0 (no smile) and 1 (a clear smile) as estimated by by Microsoft Azure Cognitive Services' computer vision API.

28. **Machine-perceived makeup** (`machine_makeup`)
    An indicator with two boolean values (True/False and True/False). The first is an indicator of machine-perceived eye makeup, the second is an indicator of machine-perceived lip makeup as estimated by Microsoft Azure Cognitive Services' computer vision API.

29. **Machine-perceived hair colors** (`machine_hair_colors`)
    A list of seven possible hair colors (black, brown, gray, other, blond, red, white) with a machine estimated confidence that this color is present in the image's hair using Microsoft Azure Cognitive Services' computer vision API, e.g., "'color': 'black', 'confidence': 0.99, 'color': 'brown', 'confidence': 0.66, 'color': 'gray', 'confidence': 0.58, 'color': 'other', 'confidence': 0.49, 'color': 'blond', 'confidence': 0.09, 'color': 'red', 'confidence': 0.05, 'color': 'white', 'confidence': 0.0." Empty cells correspond to baldness.

30. **Machine-perceived hair color** (`machine_hair_color`)
    The first color listed in `machine_hair_colors` (the color that the AI had the highest confidence was present in the hair). Empty cells correspond to baldness.

31. **Machine-perceived baldness** (`machine_bald`)
    A number between 0 and 1 describing the AI's confidence level of whether the person is bald as estimated by Microsoft Azure Cognitive Services' computer vision API.

32. **Machine-perceived glasses** (`machine_glasses`)
    Options are "NoGlasses" or "ReadingGlasses" as perceived by Microsoft Azure Cognitive Services' computer vision API. The researchers manually updated one image (awm25) in which tinted glasses were incorrectly perceived by the AI as being swimming goggles.

33. **Machine-perceived anger** (`machine_anger`)
    One of the eight machine-perceived emotions by Microsoft Azure Cognitive Services' computer vision API. The number, between 0 and 1, indicates the confidence of detection. Confidence scores are normalized, and sum to one across all eight emotion categories.

34. **Machine-perceived contempt** (`machine_contempt`)
    One of the eight machine-perceived emotions by Microsoft Azure Cognitive Services'

[5]For more details on this measure, see `https://docs.microsoft.com/en-us/azure/cognitive-services/face/concepts/face-detection`.

computer vision API. The number, between 0 and 1, indicates the confidence of detection. Confidence scores are normalized, and sum to one across all eight emotion categories.

35. **Machine-perceived disgust** (`machine_disgust`)
    One of the eight machine-perceived emotions by Microsoft Azure Cognitive Services' computer vision API. The number, between 0 and 1, indicates the confidence of detection. Confidence scores are normalized, and sum to one across all eight emotion categories.

36. **Machine-perceived fear** (`machine_fear`)
    One of the eight machine-perceived emotions by Microsoft Azure Cognitive Services' computer vision API. The number, between 0 and 1, indicates the confidence of detection. Confidence scores are normalized, and sum to one across all eight emotion categories.

37. **Machine-perceived happiness** (`machine_happiness`)
    One of the eight machine-perceived emotions by Microsoft Azure Cognitive Services' computer vision API. The number, between 0 and 1, indicates the confidence of detection. Confidence scores are normalized, and sum to one across all eight emotion categories.

38. **Machine-perceived neutral** (`machine_neutral`)
    One of the eight machine-perceived emotions by Microsoft Azure Cognitive Services' computer vision API. The number, between 0 and 1, indicates the confidence of detection. Confidence scores are normalized, and sum to one across all eight emotion categories.

39. **Machine-perceived sadness** (`machine_sadness`)
    One of the eight machine-perceived emotions by Microsoft Azure Cognitive Services' computer vision API. The number, between 0 and 1, indicates the confidence of detection. Confidence scores are normalized, and sum to one across all eight emotion categories.

40. **Machine-perceived surprise** (`machine_surprise`)
    One of the eight machine-perceived emotions by Microsoft Azure Cognitive Services' computer vision API. The number, between 0 and 1, indicates the confidence of detection. Confidence scores are normalized, and sum to one across all eight emotion categories.

41. **Machine-perceived moustache** (`machine_moustache`)
    One of the three machine-perceived facial hair categories by Microsoft Azure Cognitive Services' computer vision API. The number, between 0 and 1, indicates perceived length/thickness with 0 indicating no machine-perceivable facial hair and 1 indicating long or very thick facial hairs in a particular area.

42. **Machine-perceived beard** (`machine_beard`)
    One of the three machine-perceived facial hair categories by Microsoft Azure Cognitive

Services' computer vision API. The number, between 0 and 1, indicates perceived length/thickness with 0 indicating no machine-perceivable facial hair and 1 indicating long or very thick facial hairs in a particular area.

43. **Machine-perceived sideburns** (`machine_sideburns`)
    One of the three machine-perceived facial hair categories by Microsoft Azure Cognitive Services' computer vision API. The number, between 0 and 1, indicates perceived length/thickness with 0 indicating no machine-perceivable facial hair and 1 indicating long or very thick facial hairs in a particular area.

44. **Image size** (`image_size`)
    The size of the raw image in pixels (e.g., 160 x 200), though all photos were rated at the uniform size of 160x200.

Table A2 shows mean attractiveness ratings for each subgroup. Standard deviations are shown in parentheses.

Table A2: Average Perceived Attractiveness and Variance

		Racial Group			**Gender**		
		Black	Latino	White	Female	Male	**Total**
Professional Group	Academic	4.4 (1.0)	4.5 (1.4)	4.3 (1.2)	5.0 (1.3)	4.0 (1.0)	4.3 (1.2)
	Grad Student	4.9 (0.8)	5.1 (1.2)	4.7 (1.0)	5.3 (1.0)	4.2 (0.7)	4.9 (1.0)
	Professional	4.4 (1.2)	4.2 (1.2)	5.0 (1.4)	5.3 (1.2)	4.0 (1.1)	4.7 (1.3)
	Working Class	3.7 (0.9)	3.9 (0.9)	4.1 (1.4)	4.4 (1.4)	3.6 (0.8)	3.9 (1.2)
Gender	Female	4.8 (1.0)	5.2 (1.4)	5.1 (1.3)			5.1 (1.2)
	Male	4.2 (0.9)	3.9 (0.6)	3.9 (1.0)			4.0 (0.9)
	Total	4.5 (1.0)	4.6 (1.3)	4.5 (1.3)	5.1 (1.2)	4.0 (0.9)	

E Applications: Design and Results

We envision that the images in the RIPER database will be quite useful for a number of social science experimental designs. In particular, we believe they could be used in conjoint design experiments, which have grown common in political science and have long been popular in psychology and related fields (Bansak et al., 2021; Hainmueller, Hopkins, and Yamamoto, 2014), for experimental designs featuring campaign advertisements (e.g. Doherty and Adler, 2020; Reny, Valenzuela, and Collingwood, 2020), experiments on social media interactions (e.g. Munger, 2017), or for the more general study of race and gender (e.g. Abrajano, Elmendorf, and Quinn, 2018; Clayton, O'Brien, and Piscopo, 2019; Stephens-Dougan, 2021).

To demonstrate the potential use of the images in RIPER as primes for experiments, we used the images in our database to replicate the design of two previously-used social science experiments. Specifically, we use these images to assess the impact of attractiveness on the favorability ratings of political candidates and the impact of race on judicial sentencing decisions. Replicating the experimental design used in existing studies in which attractiveness and race play established roles serves to demonstrate the value of RIPER as a useful tool for experimental social science research. We note, however, that our results contrast with some of the findings from previous research using similar designs. This could be due to several reasons, among which are the different subject pool that we used, the amount of time since the initial studies and the changed nature of racial priming in contemporary politics and policy (Valentino, Neuner, and Vandenbroek, 2018), or simply a file-drawer problem suppressing the publication of contrasting findings. Beyond replicating the methodology of these previous studies, we also extend their designs to include additional racial groups and across numerous occupational groups.

We employed two common experimental paradigms as applications of our database of original images. First, we replicate the design of the experiment used by Lawson et al. (2010) and Todorov et al. (2005), among others, to assess the impact of attractiveness on the favorability of political candidates. Second, we conduct an experiment in which we ask respondents to make a sentencing decision on a court case for a criminal who has been pronounced guilty of vehicular homicide from a DUI offense. This application build on an experimental design used since the 1960s in research in psychology and other disciplines showing the impact of race and appearance on jury sentencing decisions (e.g. Blair, Judd, and Chapleau, 2004; Eberhardt et al., 2006; Landy and Aronson, 1969).

For the two applications, we recruited 2,052 respondents from MTurk. The experimental subjects participated in both experiments, which we randomly varied in their order. The sentencing experiment proceeded through three phases: (1) an introduction phase, (2) a description of the offense, mitigating circumstances, and aggravating circumstances for the fictional defendant Chris Sanders, and (3) an evaluation phase in which respondents reported what they thought the appropriate prison sentence was for the defendant. The candidate evaluation experiment was shorter, following Lawson et al. (2010) and Lenz and Lawson (2011), and simply presented respondents with a photo and brief information on the professional and family background of a hypothetical candidate, and then asked how likely respondents would be to vote for the candidate. Each respondent rated one candidate and sentenced one defendant.

For the sentencing experiment, we varied the photo presented to respondents of the

hypothetical defendant using a 3 (race: Latino, black, or white) x 3 (attractiveness: low, medium, or high) factorial design. A total of 9 different faces were used, which we assigned to low, medium, and high attractiveness conditions by the average rating of the faces in the database-building section of the project. The faces we used as primes, along with their average attractiveness ratings are shows in figure A1. We used only images of people whose race was accurately and reliably perceived by raters in the database-building stage. We asked respondents the sentence that they thought was most appropriate for the defendant along a scale from 1 to 25 years in prison.

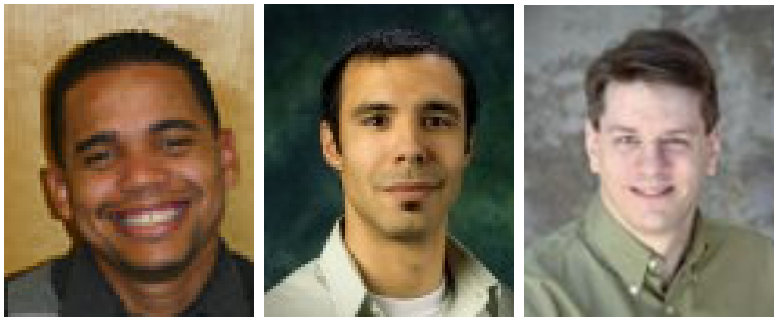
For the candidate evaluation experiment, we varied the photo of the hypothetical candidate using a 2 (race: black, or white) x 2 (gender: male or female) x 3 (attractiveness: low, medium, or high) factorial design. The face primes used are shown in Figure A2. As in the sentencing experiment, we chose photos that were of different races and attractiveness levels as determined in the database-building stage. We asked respondents how likely they were to support the candidate on a seven-point scale from extremely unlikely to extremely likely.

Figure A1: Sentencing Primes (mean attractiveness)

Low Attractiveness

(3.86) (3.79) (3.78)

Medium Attractiveness



(4.37) (4.35) (4.40)

High Attractiveness

(4.75) (4.82) (4.81)

Figure A2: Candidate Primes (mean attractiveness)

Low Attractiveness

(2.46) (3.80) (3.12) (2.72)

Medium Attractiveness

(4.58) (5.41) (4.56) (4.11)

High Attractiveness

(5.38) (7.16) (6.55) (5.66)

Results

For each of our two application experiments, we conduct an analysis of the mean dependent measure for all respondents in each treatment condition. Tables A3 and A5 show the mean prison sentence assigned or favorability rating, respectively, with the standard deviations within each condition in parentheses.

Sentencing Decisions

The results of using faces from the RIPER database contrast with experimental findings on the effect of race and attractiveness in criminal justice. We find that attractiveness has a negative but not significant impact on the sentence assigned by respondents for white defendants, but no such effect among black or Latino defendants. Our results also indicate a main effect of race, with white defendants assigned longer prison sentences, much as others have demonstrated. The interactions between race and attractiveness are all statistically insignificant. These results suggest that theories about racial priming and attractiveness may have evolved over the years since early studies on the topics were published (Valentino, Neuner, and Vandenbroek, 2018).

Table A3: Mean prison sentences assigned (standard deviations)

	Attractiveness:			
	Low	Medium	High	**Average**
Black	14.1 (7.6)	14.6 (7.9)	15.2 (7.6)	14.6 (7.7)
Latino	14.9 (7.9)	14.4 (7.4)	15.4 (7.3)	14.9 (7.5)
White	15.2 (7.6)	15.1 (7.8)	14.7 (7.8)	15.0 (7.7)

Candidate Evaluations

Our results for the candidate evaluation experiment are much in line with previous experimental results. Just as Lenz and Lawson (2011); Lawson et al. (2010) and Todorov et al. (2005) find, more attractive candidates are viewed more favorably by voters. We find a significant and positive impact of attractiveness on reported vote likelihood: a change in attractiveness from the "low" category to the "high" category caused a statistically significant increase in the likelihood a respondent would vote for the candidate by 1.5% on average. However, this pattern only holds among our white candidates. Among black candidates, the "medium" attractiveness candidates were preferred more than the "high" attractiveness candidates for both our male and female candidate primes. This effect also differed by the gender of the candidate: for white female candidates, respondents reported being 3.9% more likely to vote for them if they were one category more attractive whereas for white men, the effect was much more modest. Reading across the columns of Table A5, an interactive and nonlinear effect of attractiveness is evident. While attractiveness produced a monotonic increase in favorability among white candidates, this effect is not true among black candidates. For both black male and black female candidates, the highest level of attractiveness resulted

in less favorable ratings from respondents than the medium attractiveness condition. As with our sentencing decisions experiment, we note that these results contrast with previous research on the topic of race and attractiveness in candidate choice. However, this could be due to a different subject population, changes in the nature of stereotypes in the intervening years since previous research was published, a file-drawer problem resulting in publication of only significant results on these dynamics, or any number of other factors. Given the focus of this project on the image database rather than the results of these applications, however, we hesitate to extrapolate further about the reasons behind the differences in our results from other published research.

Table A4: Mean candidate favorability (standard deviations, N=2048)

	Attractiveness:		
	Low	Medium	High
Black Male	0.634 (0.190)	0.651 (0.198)	0.628 (0.208)
Black Female	0.632 (0.208)	0.648 (0.207)	0.639 (0.221)
White Male	0.564 (0.203)	0.588 (0.164)	0.592 (0.195)
White Female	0.557 (0.183)	0.627 (0.182)	0.642 (0.174)

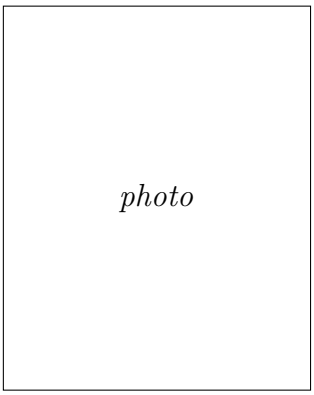
Table A5: Mean candidate favorability, white respondents only (standard deviations, N=1499)

	Attractiveness:		
	Low	Medium	High
Black Male	0.628 (0.193)	0.644 (0.194)	0.618 (0.209)
Black Female	0.616 (0.217)	0.642 (0.212)	0.628 (0.219)
White Male	0.562 (0.196)	0.589 (0.204)	0.596 (0.164)
White Female	0.550 (0.175)	0.624 (0.184)	0.6395 (0.172)

F Applications: Protocol and Stimuli

Candidate Evaluation

We are interested in the way that people perceive political candidates. We would like to know your opinions on the following candidate for office. Please just give your first impression.



Background on the candidate: [William/Paula] Forrester is a lawyer who practices in the area of criminal defense. [He/She] has been recognized for [his/her] work in the district where [he/she] is running for office, and is running on a platform focused on building local economic strength and keeping families in the district safe. [He/She] is married with two children.

How likely would you say that you are to support this candidate?

- Extremely likely
- Very likely
- Somewhat likely
- Neither likely nor unlikely
- Somewhat unlikely
- Very unlikely
- Extremely unlikely

Sentencing Decision

Introduction text:
We are interested in studying the manner in which people judge various criminal offenses. On the next page, you will read a brief account of a criminal offense. Please read this account carefully. When you have finished reading the case account, you will be asked to give your

personal opinion concerning the case.

That is, we want you to sentence the defendant described in the case account to a specific number of years of imprisonment. Remember that we are interested in your personal opinion, so please give your own personal judgment and not how you feel others might react to the case or how you feel you should react to it.

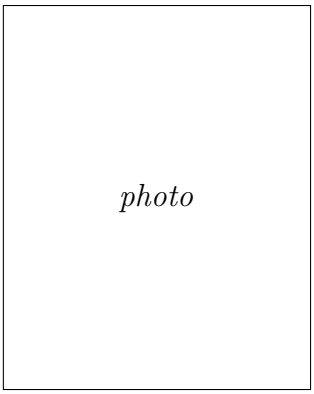
Case description:
Chris Sanders was driving home from an annual Christmas office party on the evening of December 24 when his automobile struck and killed a pedestrian by the name of Martin Lowe. The circumstances leading to this event were as follows:

The employees of the insurance office where Sanders worked began to party at around 3:00pm on the afternoon of the 24th. By 5:00pm some people were already leaving for home, although many continued to drink and socialize. Sanders, who by this time had had several drinks, was offered a lift home by a friend who did not drink and who suggested that Sanders leave his car at the office and pick it up when he was in "better shape." Sanders declined the offer, claiming that he was "stone sober" and would manage fine. By the time Sander had finished another drink, the party was beginning to break up.

Sanders left the office building and walked to the garage where he had parked his car, a four-door 2003 Chevrolet. It had just started to snow. He wished the garage attendant a Merry Christmas and pulled out into the street. Traffic was very heavy at the time. Sanders was six blocks from the garage when he was stopped by a police officer for reckless driving. It was quite apparent to the office that Sanders had been drinking, but rather than give him a ticket on Christmas Eve, he was that he would let Sanders off if he would promise to leave his car and take a taxi. Sanders agreed. The officer hailed a taxi and Sanders got into it.

The minute the taxi had turned a corner, however, Sanders told the driver to pull over to the curb and let him out. Sanders paid the driver and started back to where he had parked his own car. Upon reaching his car he proceeded to start it up and drove off. He had driven four blocks from the street where the police officer had stopped him when he ran a red light and struck Lowe, who was crossing the street. Sanders immediately stopped the car. Lowe died a few minutes later on the way to the hospital. It was later ascertained that internal hemorrhaging was the cause of death. Sanders was apprehended and charged with negligent homicide. The police medical examiner's report indicated that Sanders' estimated blood alcohol consumption was between 0.25 and 0.30 at the time of the accident. The legal limit under which it is lawful to drive is 0.08 in the state where this event happened.

About the defendant:



Sanders is employed in the area. He went to the office party in the insurance firm headquarters shortly after the party had begun. After the party Sanders was headed in the direction of home. When the incident occurred, Sanders was slightly shaken up by the impact, but suffered no major injuries. His traffic record shows he has received three traffic tickets in the past five years, two of which were moving violations. Sanders, who had stopped his car at the scene of the accident, was apprehended and charged with negligent automobile homicide, a crime which in the state is punishable by imprisonment of one to twenty-five years.

Sentencing decision:
In the next section, we would like to know what sentence, or length of punishment, you think is appropriate for the criminal case described on the previous screen.

Negligent automobile homicides similar to the one described previously are punishable by a sentence ranging from 1 to 25 years in prison.

How long of a sentence do you think the defendant should receive? [Slider: 1 to 25]

References

Abrajano, Marisa A, Christopher S Elmendorf, and Kevin M Quinn. 2018. "Labels vs. Pictures: Treatment-Mode Effects in Experiments About Discrimination." *Political Analysis* 26(1): 20–33.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins, Teppei Yamamoto, James N Druckman, and Donald P Green. 2021. "Conjoint Survey Experiments." *Advances in Experimental Political Science* 19.

Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3): 351–368.

Blair, Irene V, Charles M Judd, and Kristine M Chapleau. 2004. "The Influence of Afrocentric Facial Features in Criminal Sentencing." *Psychological Science* 15(10): 674–679.

Clayton, Amanda, Diana Z O'Brien, and Jennifer M Piscopo. 2019. "All Male Panels? Representation and Democratic Legitimacy." *American Journal of Political Science* 63(1): 113–129.

Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7(3): 613–628.

Doherty, David, and E Scott Adler. 2020. "Campaign Mailers and Intent to Turnout: Do Similar Field and Survey Experiments Yield the Same Conclusions?" *Journal of Experimental Political Science* 7(2): 150–155.

Eberhardt, Jennifer L, Paul G Davies, Valerie J Purdie-Vaughns, and Sheri Lynn Johnson. 2006. "Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes." *Psychological Science* 17(5): 383–386.

Hainmueller, Jens, Daniel J Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1): 1–30.

Landy, David, and Elliot Aronson. 1969. "The Influence of the Character of the Criminal and His Victim on the Decisions of Simulated Jurors." *Journal of Experimental Social Psychology* 5: 141–152.

Lawson, Chappell, Gabriel S. Lenz, Michael Myers, and Andy Baker. 2010. "Candidate Appearance, Electability, and Political Institutions: Findings from Two Studies of Candidate Appearance." *World Politics* 62(4): 561–593.

Lenz, Gabriel S., and Chappell Lawson. 2011. "Looking the Part: Television Leads Less Informed Citizens to Vote Based on Candidates' Appearance." *American Journal of Political Science* 55(3): 574–589.

Munger, Kevin. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39(3): 629–649.

Reny, Tyler T, Ali A Valenzuela, and Loren Collingwood. 2020. ""No, You're Playing the Race Card": Testing the Effects of Anti-Black, Anti-Latino, and Anti-Immigrant Appeals in the Post-Obama Era." *Political Psychology* 41(2): 283–302.

Stephens-Dougan, LaFleur. 2021. "The Persistence of Racial Cues and Appeals in American Elections." *Annual Review of Political Science* 24: 301–320.

Todorov, Alexander, Anesu N Mandisodza, Amir Goren, and Crystal C Hall. 2005. "Inferences of Competence from Faces Predict Election Outcomes." *Science* 308(5728): 1623–1626.

Valentino, Nicholas A, Fabian G Neuner, and L Matthew Vandenbroek. 2018. "The Changing Norms of Racial Political Rhetoric and the End of Racial Priming." *Journal of Politics* 80(3): 757–771.