When is Discrimination Unfair?

Peter Kuhn

Trevor Osaki

Appendix (for online publication)

February 16, 2023

Appendix 1: Survey Design

A1.1 Instructions and Questions

This section reproduces the instructions and questions that were encountered by a participant who was allocated to the TB (Tastes, Black) and SB (Statistical, Black) treatment combinations in Stages 1 and 2 respectively. White treatments were identical to the Black treatments with the races of the discriminator and discriminatee reversed. Less and more justifiable forms of discrimination were administered in random order within a Stage. Items in [square brackets] were not seen by the participants.

[Overall Introduction]

In this survey, you will be asked to read and react to four hypothetical scenarios, or vignettes that happen in a workplace. We will also ask you to explain one of your choices and collect some background information about you.

The scenarios you'll evaluate have been randomly selected from a larger variety of situations we are asking many people about. These situations describe different types of people interacting in different ways.

Some of these scenarios may seem realistic to you; others may seem unrealistic. In all cases you will have only very limited information about what happened.

Regardless of how likely you think these situations might be, and despite the limited information, we ask that you please give us your reaction to them if they were to happen, based on the information that has been provided.

[Stage 1 Introduction]
Please read the following two hypothetical scenarios carefully. They are similar in many respects, but they differ in a few ways. **To help you see the differences**, we have underlined them. After you read each scenario, we will ask you for your reaction to it.

Situation 1 [Tastes, Black, *less* justifiable *(based on own tastes)***]:**

Michael, who is White, is about to hire his first customer representative for his business after a few years of carrying that role alongside his managerial duties. He has interacted with a number of Black people during his education and work experience. While all of his interactions with Black people have been polite and professional, he just didn't enjoy interacting with them.

For his new hire, Michael has to choose between two applicants whose resumes, interviews and references are all of equal quality, one of whom is Black and one who is White. Michael decides to hire the White worker in order to avoid interacting with a Black employee.

Given the information provided in the preceding scenario, please indicate the extent to which you thought that Michael's hiring decision was fair:

[Choose one from: 1-very unfair, 2-unfair, 3-somewhat unfair, 4-neither fair nor unfair, 5-somewhat fair, 6-fair, 7-very fair].

Situation 2 (Tastes, Black, *more* justifiable *(based on others' tastes)*]**:**

Michael, who is White, is about to hire his first customer representative for his business after a few years of carrying that role alongside his managerial duties. He has conducted focus groups with a substantial share of the people who frequent his business. Many of these customers tell Michael that they do not like interacting with Black people and would be hesitant about continuing to support his business if he employed them. Michael himself is just as happy to interact with Black workers as with workers of other races.

For his new hire, Michael has to choose between two applicants whose resumes, interviews and references are all of equal quality, one of whom is Black and one who is White. Michael decides to hire the White worker, in order to avoid losing sales to customers who do not want to interact with Black representatives.

Given the information provided in the preceding scenario, please indicate the extent to which you thought Michael's hiring decision was fair.

[Choose one from: 1-very unfair, 2-unfair, 3-somewhat unfair, 4-neither fair nor unfair, 5-somewhat fair, 6-fair, 7-very fair].

[Stage 2 Introduction]

Please read the following two scenarios carefully. As a result of random assignment, the **types of people** involved and their actions **may or may not** change from the last two scenarios.

Like the first two scenarios, the next two scenarios are quite similar to each other. **To help you see the differences**, we have underlined them. After you read each scenario, we will ask you for your reaction to it.

Situation 1 [Black, Statistical, *less* justifiable *(based on hearsay)*]:

Michael, who is White, is about to hire his first customer representative for his business after a few years of carrying that role alongside his managerial duties. He has discussed his business plans with a neighbor. This neighbor says he once met a business owner who had trouble with some Black employees. Problems included unexcused absenteeism, being late for work, and a lack of attention to detail on the job.

For his new hire, Michael has to choose between two applicants whose resumes, interviews and references are all of equal quality, one of whom is Black and one who is White. Michael decides to hire the White worker based on a brief conversation he had with his neighbor about problems with Black workers.

Situation 2 [Black, Statistical, *more* justifiable *(based on higher quality information)*]:

Michael, who is White, is about to hire his first customer representative for his business after a few years of carrying that role alongside his managerial duties. He has discussed his business plans with a large and experienced network of local business owners who frequently hire customer representatives. They tell Michael that they have had trouble with a large share of their Black representatives, and they show Michael some reliable statistics from their businesses that verify these claims. Problems included unexcused absenteeism, being late for work, and a lack of attention to detail on the job.

For his new hire, Michael has to choose between two applicants whose resumes, interviews and references are all of equal quality, one of whom is Black and one who is White. Michael decides to hire the White worker, based on the information and statistics about local Black workers that he got from experienced local business owners.

[Stage 3/Follow-up Introduction]

Recall the scenario that you just evaluated, in which [brief description of second scenario encountered in Stage 1]. You thought that Michael's hiring decision was [very unfair/unfair/somewhat unfair/neither fair nor unfair/somewhat fair/fair/very fair]. In 50 words or less, please explain your response.

If you would like to skip this question, please type: "Prefer not to answer."

1. This question refers to the final vignette encountered. **[Open-ended]**.

You thought that Michael's hiring decision was [very unfair / unfair / somewhat unfair / neither fair nor unfair / somewhat fair / fair / very fair]. In 50 words or less, please explain your response.

2. Please consider the following question without referring to any of the previous survey items, and then select the rating that best corresponds to your answer:

All in all, in the United States, how would you compare the economic opportunities available to Black and White people?

[Choose one from:]

1-Black people have much less opportunity than White people,

2-Black people have less opportunity than White people,

3-Black people have a little less opportunity than White people,

4-Black and White people have roughly equal opportunities,

5-Black people have a little more opportunity than White people,

6-Black people have more opportunity than White people,

7-Black people have much more opportunity than White people]

[Background Questions Introduction]

Please answer the following background questions.

1. Please indicate your gender.

- Male
- Female
- Other/decline to state

2. Please indicate your age.

- 18-28
- 25-34
- 35-44
- 45-54
- 55-64
- 65-74
- 75-84
- 85 and older

3. Please indicate the highest level of education you have completed.

- Primary school or below (grades 1-8)
- High School (grades 9-12)
- Some College (includes two-year college degrees)
- Four-year College or University Degree
- Higher Degree (e.g., MD, MBA, Master's, PhD)

4. Please select the category that best describes your race.

- Hispanic, Latino, or Spanish origin
- White
- Black or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Other Pacific Islander
- Other

5. What is your U.S. political party preference?

- Democrat
- Republican
- Independent or no party affiliation
- Other

6. Which of these best describes your political views?

- Extremely liberal
- Liberal
- Slightly liberal
- Moderate
- Slightly Conservative
- Conservative
- Extremely Conservative

[Final instructions]

Here is your ID: **####**

To receive your payment for participating, click "Accept HIT" in the MTurk window, enter this ID number, and then click "submit."

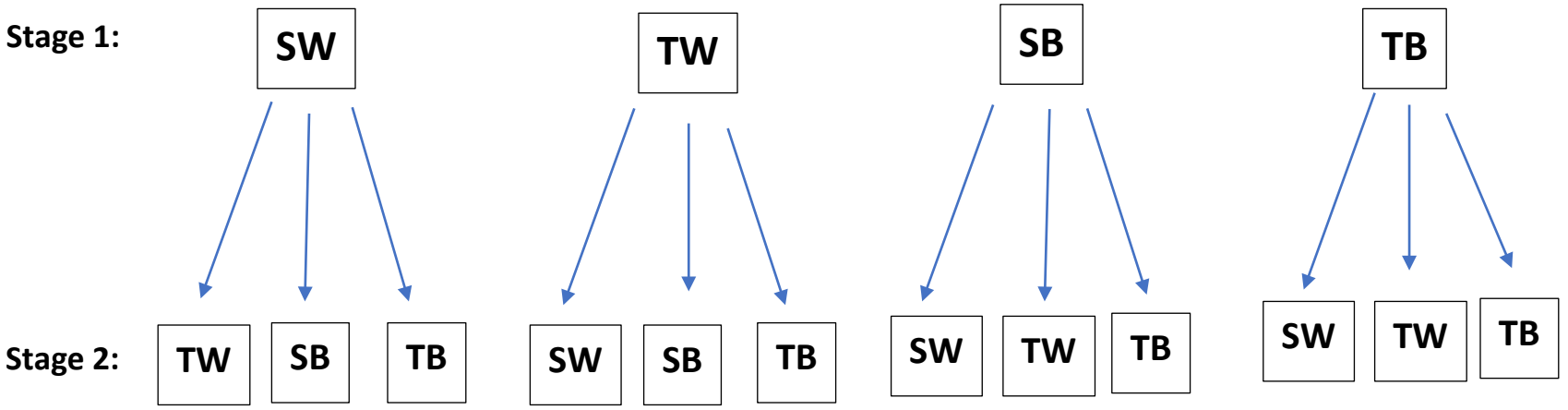
Please do not exit the survey from this page. You must click on the "next button" to reach the "end of survey" page so that your responses are recorded. This button will appear in a few seconds.

A1.2 Randomization

As illustrated in Figure A1.2.1, subjects were randomly assigned to one of four treatment combinations in Stage 1 of the Survey. In Stage 2, subjects were randomly re-assigned to one of the three treatment combinations they had not encountered in Stage 1. Within each Stage, the more- versus less-justifiable versions of the scenarios for that treatment combination were administered in random order.

Thus, two thirds of the subjects experienced a change in the Statistical / Tastes treatment, and two thirds experience a change in the *race* treatment. The discriminator's name (Michael or Andrew) was randomly assigned in Stage 1, then switched for all respondents in Stage 2.

Figure A1.2.1 Randomization in Stages 1 and 2



Notes:

T = Tastes; S = Statistical; B = Black; W = White (race refers to the *discriminatee*)

In each Stage, respondents were assigned across treatments with equal probability.

Notes: This figure illustrates how the survey treatments are randomized between Stages 1 and 2. SW, TW, SB, and TB refer to combinations of *motivation* and *race* treatments that are allocated to a Stage. For example, SW refers to a set of vignettes illustrating statistical discrimination where the discriminatee is White. Respondents were assigned one of (SW, TW, SB, and TB) with equal probability in Stage 1. In Stage 2, they were assigned a treatment combination they did not encounter in Stage 1.

Appendix 2: Representativeness

Table A2.1 shows the mean demographic characteristics of our MTurk sample in column (1). Column (2) contains means of the same characteristics for adults in the 2019 American Community Survey (ACS), a nationally representative survey sample, for comparison. As is well known, MTurkers are more male, better educated, and much more likely to be between 25 and 44 years of age than U.S. adults in general. MTurkers are also slightly more likely to be White and Black, and less likely to belong to other racial groups than the U.S. population.

Table A2.2 shows the mean shares of respondents by political orientation of our MTurk respondents in column (1). Column (2) contains these means from the General Social Survey (GSS), another nationally representative survey sample.[1] Overall, Table A2.2 suggests that MTurk respondents differ from the GSS in two main ways: First, compared to the GSS a smaller share of MTurk respondents choose the middle three categories: 'moderate' or 'slightly' liberal / conservative, while MTurkers are also more likely to locate in the two 'extreme' categories. In this sense, MTurkers are politically more extreme than GSS respondents. It is possible, however, that some of this is caused by a difference in phrasing of the middle category between the two surveys. Second, almost identical shares of MTurkers and GSS respondents choose some degree of conservative leaning (ranging from slight to extreme), but many more MTurkers choose some liberal leaning (47.3 versus 30.2 percent). Thus, on average, MTurkers are more liberal than the U.S. population as a whole.

Tables A2.3 and A2.4 compare the geographical distribution of our MTurk sample obtained from the approximate geocoordinates of respondents recorded by the survey software to the distribution of the adult ACS population by Census regions/subregions and across states with populations of 5 million or more. (MTurk sample shares become very imprecise in smaller states). While MTurkers are slightly more likely to live in the Northeast and West, they are widely represented across all the larger states, with no clear pattern in over- versus under-representation.

Finally, Figure A2.1 shows Google search trends for "Black Lives Matter", "racism" and "discrimination" during the period surrounding our survey. It shows that the high level of public concern surrounding these issues associated with the killing of George Floyd had essentially dissipated by the time our survey was in the field.

[1] Since the ACS does not collect information on political opinions or affiliations, we are forced to use the GSS (with its much smaller sample size) to assess the representativeness of our population. Our political party preference question is not comparable to the GSS's, but our political leaning question is almost identical to the GSS's (see Table A2.2 for details).

Table A2.1: Demographic Composition of MTurk Sample versus the American Community Survey (ACS)

CHARACTERISTIC	MTurk Sample (1)	2019 ACS Sample (2)
Male	0.600	0.485
Female	0.400	0.515
White respondent	0.780	0.713
Black respondent	0.115	0.090
Asian respondent	0.042	0.084
Hispanic respondent	0.037	0.020
American Indigenous respondent	0.009	0.010
Pacific Islander respondent	0.005	0.003
Other race respondent	0.011	0.080
Age 18-24	0.037	0.103
Age 25-34	0.435	0.152
Age 35-44	0.294	0.148
Age 45-54	0.146	0.156
Age 55-64	0.061	0.181
Age 65 and over	0.026	0.291
High School or less	0.098	0.362
2-year or some college	0.196	0.307
4-year college or university	0.519	0.203
Higher degree	0.187	0.128
Observations	642	846,557

Notes: Column 1 contains the percentage of respondents across various demographic characteristics within the MTurk sample. Column 2 contains these percentages for the 2019 American Community Survey (ACS) sample for comparison. The racial categories in our ACS data use the mutually exclusive categories derived by Center for Economic and Policy Research (CEPR) (variable *wbhapo*), which match our own survey question.

Table A2.2: Composition of MTurk Sample versus the General Social Survey (GSS), by Political Leaning

CHARACTERISTIC	MTurk Sample (1)	GSS Sample (2)
Extremely conservative	0.101	0.051
Conservative	0.164	0.168
Slightly conservative	0.092	0.146
Moderate	0.170	0.332
Slightly liberal	0.095	0.121
Liberal	0.274	0.132
Extremely liberal	0.104	0.049
Observations	642	1,776

Notes: Column 1 contains the percentage of respondents by political leaning while Column 2 contains that of the 2020 GSS. Our political party preference question is not comparable to the GSS. The only difference between our political leaning question and the GSS is in the phrasing of the middle category:

Our political leaning question asks for "political views" on this seven-point scale:

extremely liberal; liberal; slightly liberal
moderate;
slightly conservative; conservative; extremely conservative

The GSS political leaning question ask for "political views" on this seven-point scale:

extremely liberal; liberal; slightly liberal
moderate, middle of the road;
slightly conservative; conservative; extremely conservative

Table A2.3: Composition of MTurk Sample by Census Region

CENSUS REGION	MTurk Sample (1)	2019 ACS Sample (2)
Northeast	0.238	0.178
New England	0.028	0.048
Middle Atlantic	0.210	0.130
Midwest	0.189	0.212
East North Central	0.136	0.146
West North Central	0.053	0.066
South	0.394	0.376
South Atlantic	0.251	0.201
East South Central	0.047	0.059
West South Central	0.097	0.116
West	0.179	0.234
Mountain	0.051	0.073
Pacific	0.128	0.161
Observations	642	2,599,171

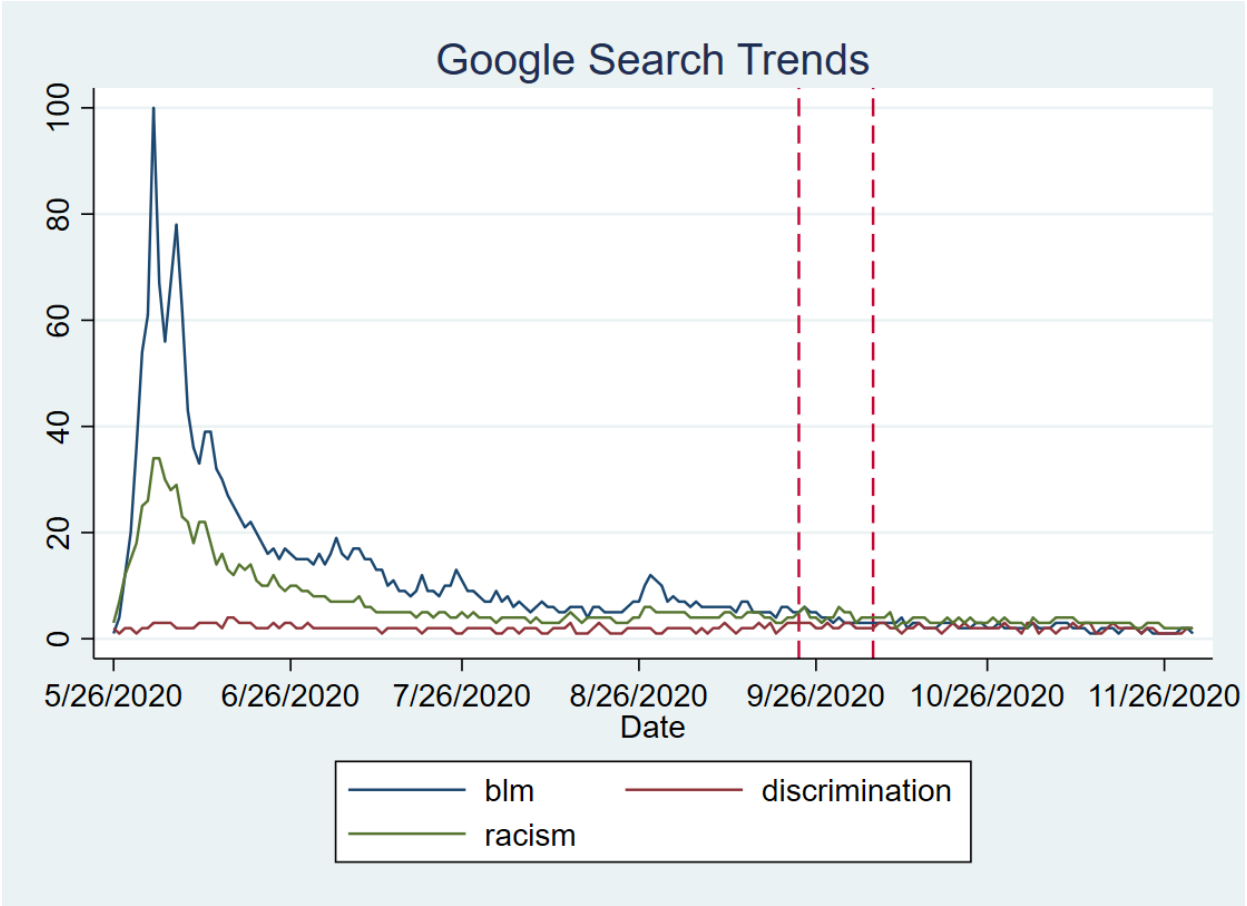
Notes: Column 1 contains the percentage of respondents across U.S. census regions and their respective divisions. Column 2 contains these percentages for the 2019 American Community Survey (ACS) sample for comparison.

Table A2.4: Composition of MTurk Sample versus ACS by U.S. State (pop. exceeds 5 million)

STATE	MTurk Sample *shares* (1)	MTurk Sample *Count* (2)	2019 ACS Sample *shares* (3)	State Pop. *in thousands* (4)
Arizona	0.023	11	0.031	5,638
California	0.119	57	0.167	30,618
Florida	0.131	63	0.094	17,248
Georgia	0.040	19	0.044	8,114
Illinois	0.060	29	0.055	9,854
Indiana	0.033	16	0.029	5,164
Massachusetts	0.013	6	0.032	5,540
Michigan	0.029	14	0.044	7,843
New Jersey	0.048	23	0.039	6,944
New York	0.158	76	0.089	15,425
North Carolina	0.038	18	0.046	8,187
Ohio	0.048	23	0.053	9,111
Pennsylvania	0.075	36	0.058	10,167
Tennessee	0.019	9	0.030	5,319
Texas	0.092	44	0.116	21,596
Virginia	0.040	19	0.037	6,675
Washington	0.035	17	0.034	5,952
Observations	480	480	1,821,247	-

Notes: Column 1 contains the percentage of respondents across U.S. states with adult populations of at least 5 million. Column 2 contains the raw number of MTurk respondents from each state. Column 3 contains the percentages for the 2019 American Community Survey (ACS) sample for comparison. Column 4 contains the 2019 state populations (in thousands) of those at least 18 years of age.

Figure A2.1: Frequency of Google Searches for BLM and Related Keywords around the Survey Date



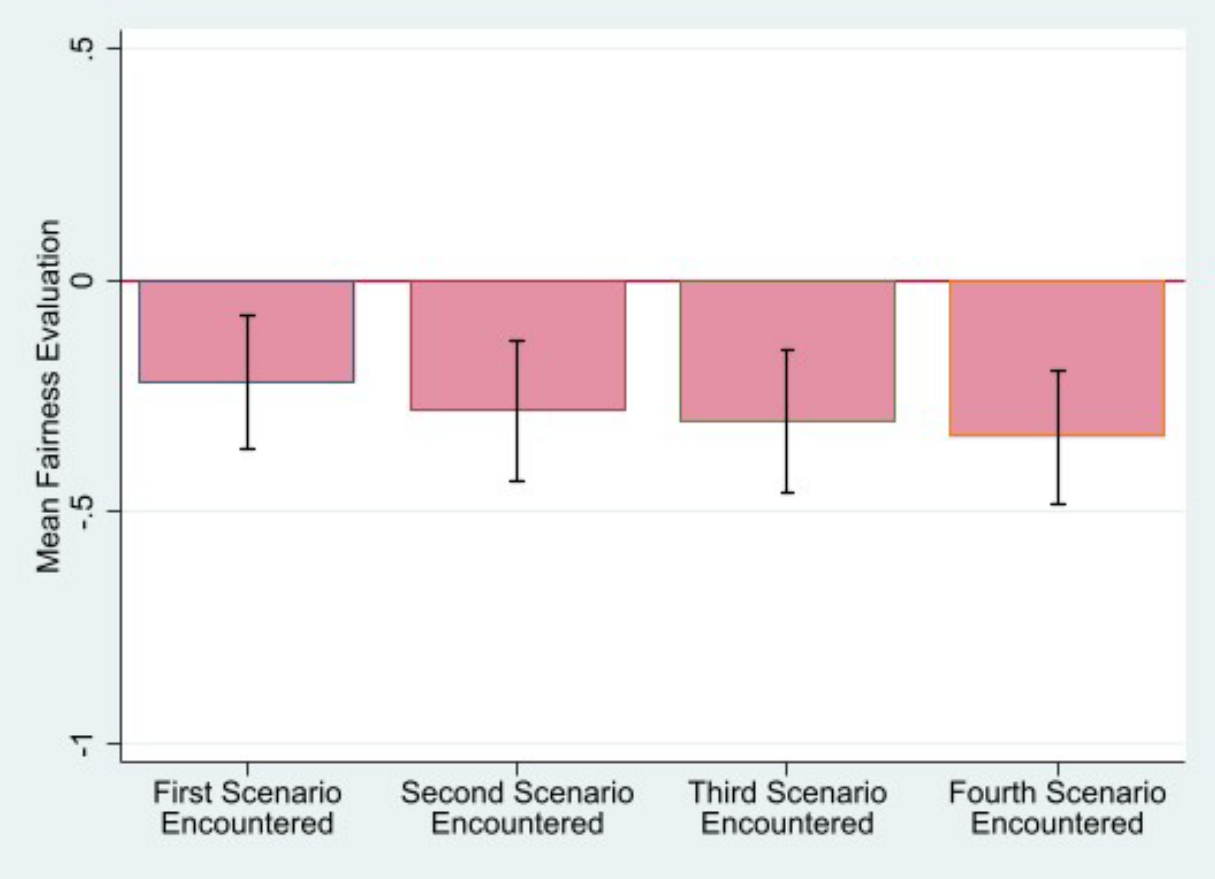
Note: This figure illustrates trends in Google searches for keywords related to three topics: "Black Lives Matter (blm), racism, and discrimination. The vertical axis represents search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. The region bounded by the two dotted lines represent the dates our survey was live on MTurk. The data on these interest values was drawn from Google Trends.

Appendix 3: Order Effects

A3.1 Pure Order Effects

Figure A3.1 shows there is no strong association between the respondents' fairness evaluations and the order of scenarios they encountered throughout the survey.

Figure A3.1.1



Notes: The *p*-values below are clustered by respondent.

- First scenario vs. second = 0.412
- Second scenario vs. third = 0.778
- Third scenario vs. fourth = 0.644
- Fourth scenario vs. first = 0.112

A3.2 Order Effects for the Taste versus Statistical Treatments

In this Section we test for whether the order in which the respondents encounter the Taste and Statistical treatments affects their fairness assessments. First, we compare the Stage 2 fairness ratings of workers who received different treatments in Stage 1. Next, we compare the within-subject fairness changes of respondents who switched from to Tastes to Statistical to the changes of respondents who switched in the other direction. Finally, we compare aggregate, within-subject, and between-subject regression estimates of the Taste treatment effect. None of these exercises reveal any treatment order effects.

A3.2.1 Stage 2 Assessments as a Function of Stage 1 Treatment

Figure A3.2.1 (a) shows that Respondents who encountered Taste-based scenarios in Stage 1 view Statistical and Taste discrimination as equally fair in Stage 2. Figure A3.2.1 (b) shows that respondents who encountered Statistical scenarios in Stage 1 also view Statistical and Taste discrimination as equally fair in Stage 2. Thus, we see no evidence of order effects.

Figure A3.2.1: Stage 2 Fairness Assessments by Stage 1 Treatment: Taste versus Statistical

***p*-values (clustered by respondent)**:

A vs B = 0.834

C vs D = 0.755

A vs C = 0.675

B vs D = 0.314

A3.2.2 Ratings Changes of Subjects Who Switched Treatments

We cannot reject that the fairness ratings changes of respondents who were switched from the Taste to the Statistical treatment between Stages 1 and 2 are equal but opposite in sign to respondents who were switched in the opposite direction. Specifically, the ratings change of Taste-Statistical switchers was -0.113 ($p = .288$); the ratings change of Statistical-Taste switchers was -0.091; ($p = .344$). A test for equality between these two changes cannot reject the null (p = .879; clustered by respondent).

A3.2.3 Comparing within-subject, between-subject and pooled estimates of the Taste treatment effect

Figure A3.2.3 presents three types of regression estimates of the Taste treatment effect. *Within-subject* estimates regress fairness on a treatment indicator (i.e., it takes on a value of "1" if the scenario illustrates taste-based discrimination) plus respondent fixed effects. *Between-subject* estimates are pure cross-section regressions using data from only the first of the four scenarios each respondent encountered. Pooled estimates include all four scenarios each person encountered, without person fixed effects. All three treatment effects are very small in magnitude and indistinguishable from zero. Tests for equality between all pairs of estimated treatment effects cannot reject the null hypothesis.

Figure A3.2.3: Comparison of Taste Treatment Effect Estimates

Notes:

1. The p-values below are clustered by respondent:
    - Between vs. Within-subject = 0.574
    - Within-subject vs. Pooled = 0.312
    - Pooled vs. Between-subject = 0.205

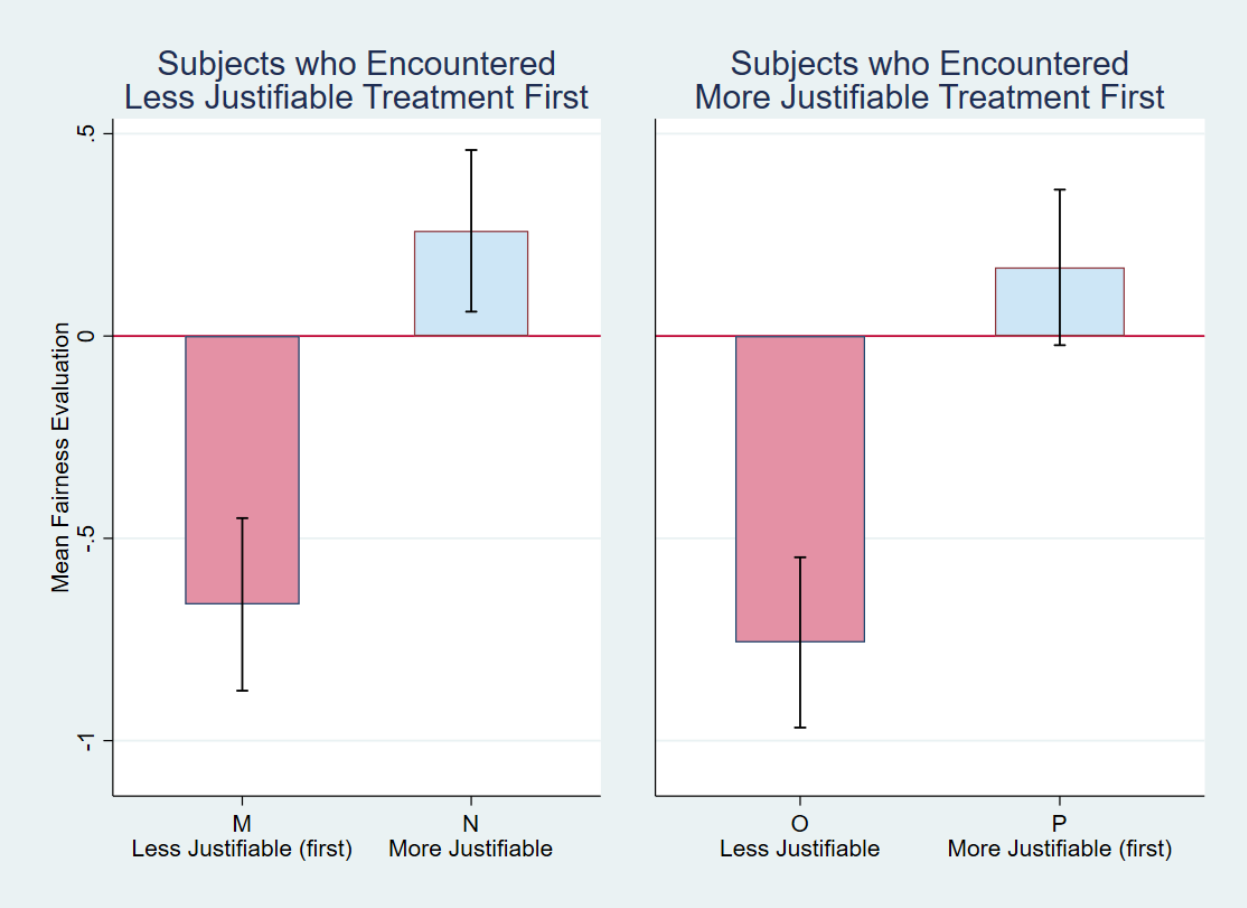
A3.3 Order Effects for the *Less* versus *More* Justifiable Treatments

In this Section we test for whether the order in which the respondents encounter the *less* versus *more* justifiable scenarios affects their fairness assessments. We focus first on the effects of *justifiability* treatment variation within Stage 1, next on variation within Stage 2, and then pool the within-Stage variation from both Stages. Finally, we compare aggregate, within-subject, and between-subject regression estimates of the *less* justifiable treatment using data from the entire survey. None of these exercises reveal any treatment order effects.

A3.3.1 *Justifiability* Treatment Variation within Stage 1

Figure A3.3.1 focuses on treatment order effects within Stage 1, and shows that respondents' fairness evaluations of the *less* and *more* justifiable treatments in the second scenario they encountered do not depend on which of those treatments they encountered in the preceding scenario. It also shows that the ratings changes of *less*- to *more-justifiability* switchers are statistically equal but opposite in sign the ratings changes of *more*- to *less-justifiability* switchers

Figure A3.3.1: Mean Fairness Ratings by the First Scenario Encountered in Stage 1



Notes: The *p*-values below are clustered by respondent.

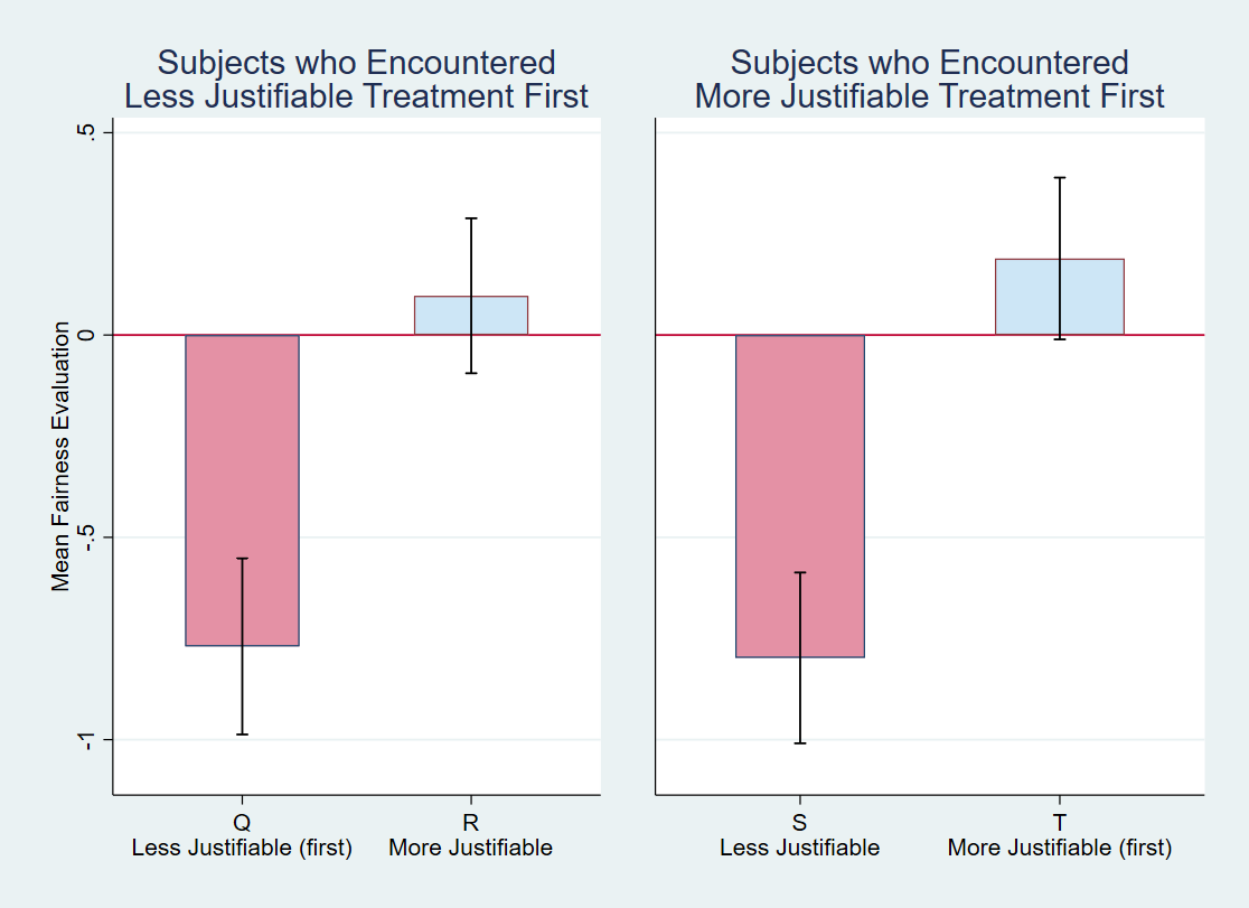
- M vs. N = 0.000
- O vs. P = 0.000
- M vs. O = 0.537
- N vs. P = 0.521

Equality test for switchers: M – N = O – P: $p = .979$

A3.3.2 *Justifiability* Treatment Variation within Stage 2

Figure A3.3.1 focuses on treatment order effects within Stage 2, and shows that respondents' fairness evaluations of the *less* and *more* justifiable treatments in the second scenario they encountered do not depend on which of those treatments they encountered in the preceding scenario. It also shows that the ratings changes of *less*- to *more-justifiability* switchers are statistically equal but opposite in sign the ratings changes of *more*- to *less-justifiability* switchers

Figure A3.3.2: Mean Fairness of Respondents by the First Scenario they Encountered in Stage 2



Notes: The *p*-values below are clustered by respondent.

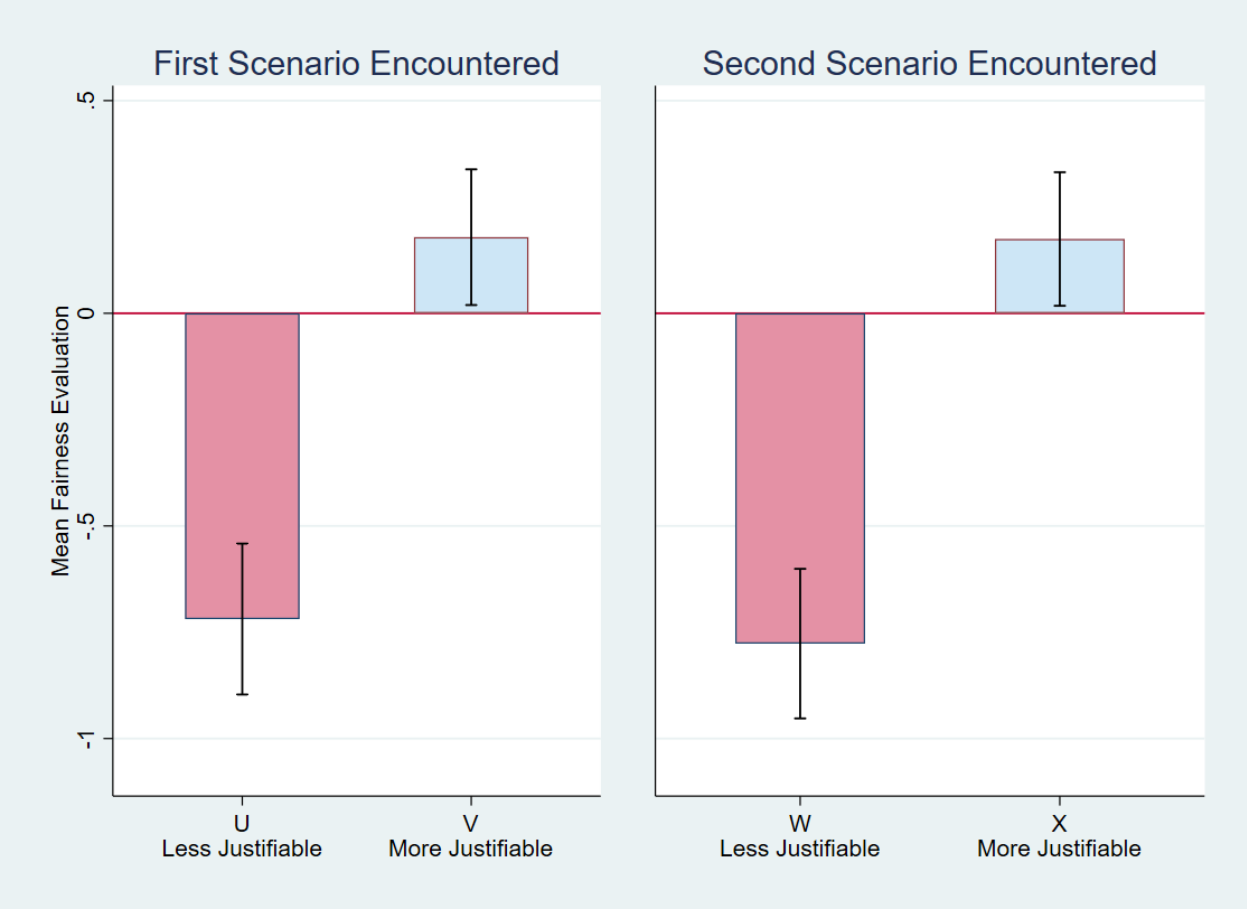
- Q vs. R = 0.000
- R vs. S = 0.000
- Q vs. S = 0.854
- R vs. T = 0.513

Equality test for switchers: Q – R = S – T: $p = .350$

A3.3.3 Pooling within-Stage *Justifiability* Treatment Variation from both Stages

Figure A3.3.3 pools data from the two Stages of our survey, and continues to find that subjects' fairness evaluations of the *less* and *more* justifiable scenarios do not depend on which one they encountered previously in the current Stage of the survey. Once again, the fairness changes of the *less-to-more* switchers are statistically equal but opposite in sign to the *more-to-less* justifiable switchers.

Figure A3.3.3: Mean Fairness of Respondents by the Scenario Ordering they Encountered, Pooling Stages 1 and 2



Notes: The *p*-values below are clustered by respondent.

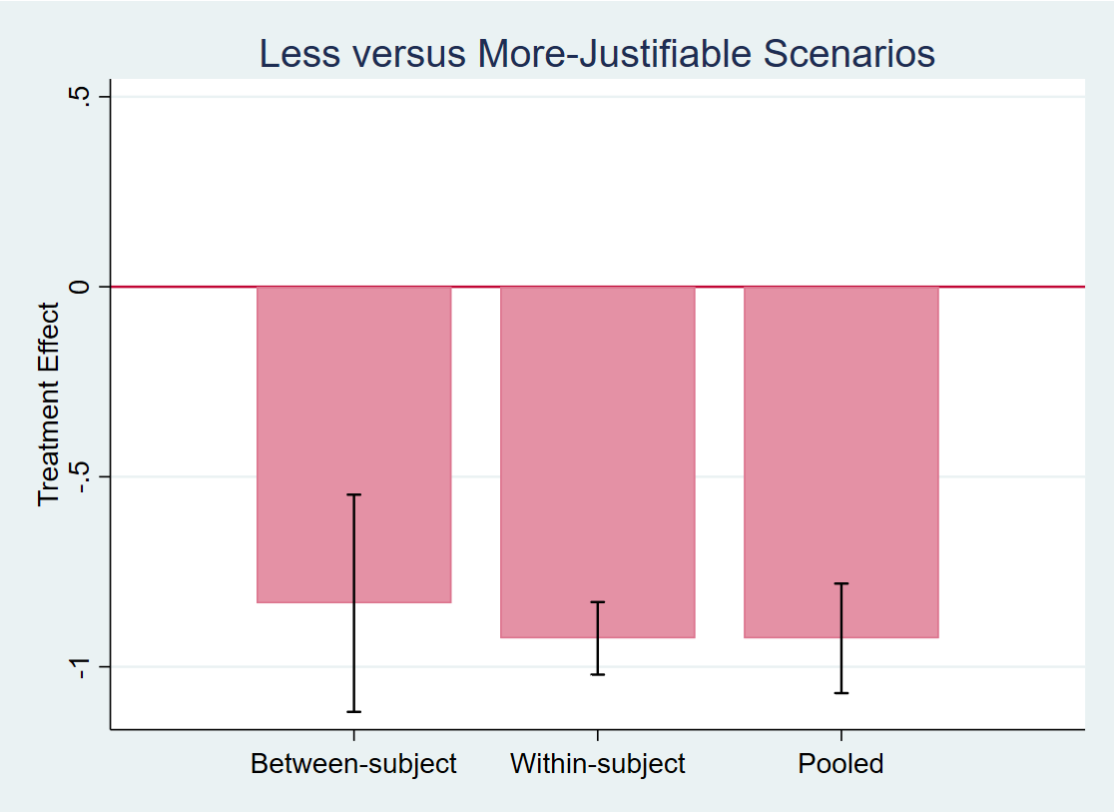
- U vs. V = 0.000
- W vs. X = 0.000
- U vs. W = 0.610
- V vs. X = 0.967

Equality test for switchers: U – V = W – X: $p = .782$

A3.3.4 Comparing within-subject, between-subject, and pooled estimates of the *less* justifiable treatment effect

Using data from all four scenarios each respondent encountered in the survey, Figure A3.3.4 compares within-subject, between-subject and pooled regression estimates of the *less* justifiable treatment on subjects' fairness assessments. All three estimates of the treatment effect are substantial in magnitude, negative, and statistically significant. In addition, all three estimates are very similar, and are statistically indistinguishable from each other.

Figure A3.3.4: Comparison of *Less* Justifiable Treatment Effect Estimates



Notes:

1. The *p*-values below are clustered by respondent.
    - Between vs. Within-subject = 0.498
    - Within-subject vs. Pooled = 1.00
    - Pooled vs. Between-subject = 0.498

2. Within-subject estimates regress fairness on a treatment indicator plus respondent fixed effects. Between-subject estimates are pure cross-section regressions using data from the first scenario each respondent encountered only. Pooled estimates include all four scenarios each person encountered, without person fixed effects.

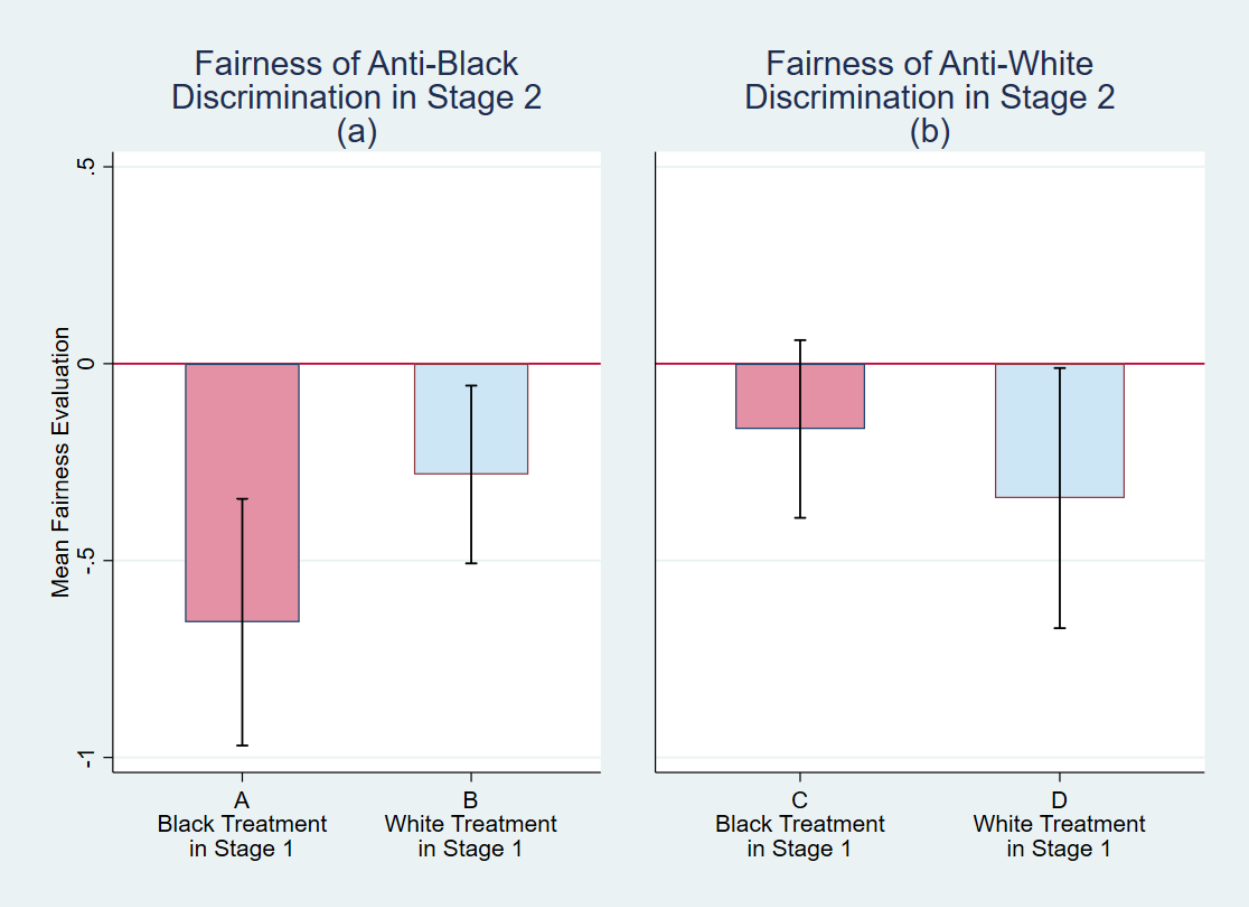
A3.4 Order Effects for the Race Treatment

In this Section we test for whether the order in which the respondents encounter a Black versus a White discriminatee affects their fairness assessments. First, we compare the Stage 2 fairness ratings of workers who received different treatments in Stage 1. Next, we compare the within-subject fairness changes of respondents who switched from to Black to White to the changes of respondents who switched in the other direction. Finally, we compare aggregate, within-subject, and between-subject regression estimates of the Black treatment effect. Overall, we find substantial evidence of a particular type of treatment order effect: Subjects who encountered the White treatment in Stage 1 were more tolerant of anti-Black discrimination in Stage 2 (compared to subjects who encountered the Black treatment in Stage 1).

A3.4.1 Stage 2 Assessments as a Function of Stage 1 Treatment

Figure A3.4.1 (a) shows subjects' Stage 2 fairness assessments, separately for subjects who encountered the Black versus White treatment in Stage 1. In contrast to the preceding results for the Statistical versus Tastes or the *less* versus *more* justifiable treatments, treatment order matters here. Specifically, subjects who encountered anti-Black discrimination in Stage 2 rated it more harshly if they also encountered it in Stage 1, compared to subjects who encountered anti-White discrimination in Stage 1.

Figure A3.4.1: Subjects' Stage 2 Fairness Assessments, by their Stage 1 *Race* Treatment



***p*-values**:

A vs B = 0.055

C vs D = 0.385

A vs C = 0.012

B vs D = 0.767

Notes: All *p*-values are clustered by respondent.

A3.4.2 Ratings Changes of Subjects Who Switched *Race* Treatments

The mean ratings change of Black-to-White switchers was 0.243 ($p = .005$); the ratings change of White-to-Black switchers was -0.381; ($p = .000$). A test for equality between these two ratings changes indicated that they are statistically distinguishable from each other ($p = .000$).

A3.4.3 Comparing within-subject, between-subject and pooled estimates of the *Race* treatment effect

Figure A3.2.3 presents three types of regression estimates of the *race* treatment effect. *Within-subject* estimates regress fairness on a treatment indicator (i.e., it takes on a value of "1" if the discriminatee is Black) plus respondent fixed effects. *Between-subject* estimates are pure cross-section regressions using data from only the first of the four scenarios each respondent encountered. Pooled estimates include all four scenarios each person encountered, without person fixed effects. The figure shows that the within-subject and pooled estimates are similar in magnitude, and they are statistically indistinguishable from each other. However, the between-subject estimate is roughly twice as large as those two estimates and statistically distinguishable from them.

Figure A3.2.3: Comparison of Black Treatment Effect Estimates

Notes: The *p*-values below are clustered by respondent:

- Between vs. Within-subject = 0.001
- Within-subject vs. Pooled = 0.647
- Pooled vs. Between-subject = 0.007

Appendix 4: Exploring the Effects of Education on Fairness Ratings

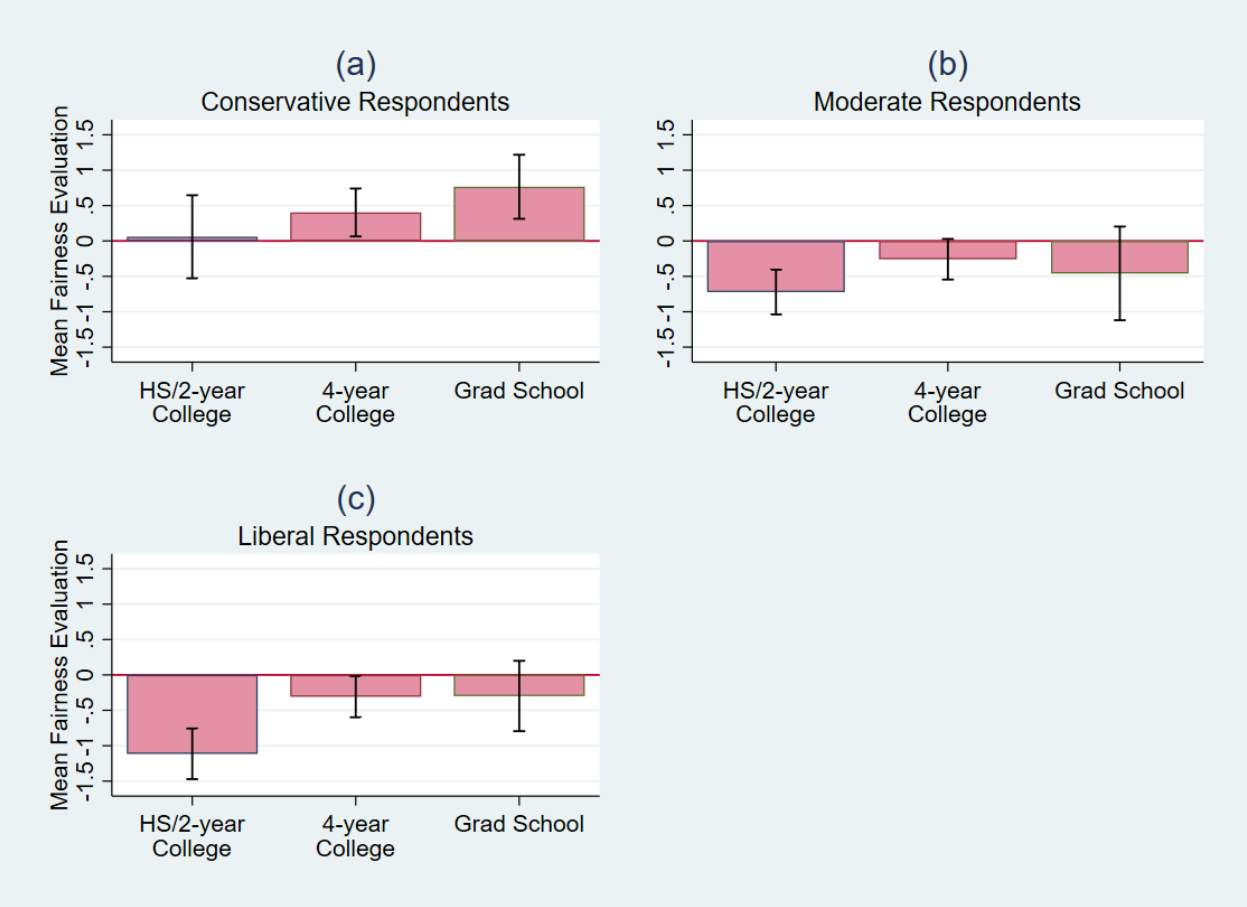
This Section explores the unexpected (to us) positive association between respondents' education and their ratings of the fairness of discriminatory actions. We show, first of all, that the positive association between education and fairness is not an artifact of political differences between the education groups. Instead, Figure A4.1 shows that education is associated with increased perceived fairness within each of our three political groups. Next, while our respondents' political leanings affect the way they respond to our Race treatment, we show that education does not have this effect: Despite being more tolerant of discriminatory acts in general, respondents of all education levels react more negatively to anti-Black and to anti-White discrimination (Figure A4.2). In fact, this discriminatee race effect is remarkably constant across education groups, despite the differences in their mean fairness assessments.

Finally, one of our main findings in the paper is that conservatives do not exhibit a discriminatee race effect, while moderates and liberals do. In Figure A4.3, we show that education differences do not account for this fact either. In fact, our that liberals exhibit a discriminatee race effect and conservatives do not is present *within all three education groups* (Figure A4.3).

Taken together, these three findings show that the positive education-fairness association is broadly distributed across political groups and experimental treatments, and does not affect how people respond to our experimental treatments. Thus, we conclude that it likely reflects different *set points* for fairness by education rather than differences in political affiliation or racial attitudes across education groups.

Figure A4.1: Mean Fairness Assessments by Education and Political Leaning

The positive association between education and fairness is not an artifact of political differences between the education groups- we see it within each of our three political groups:

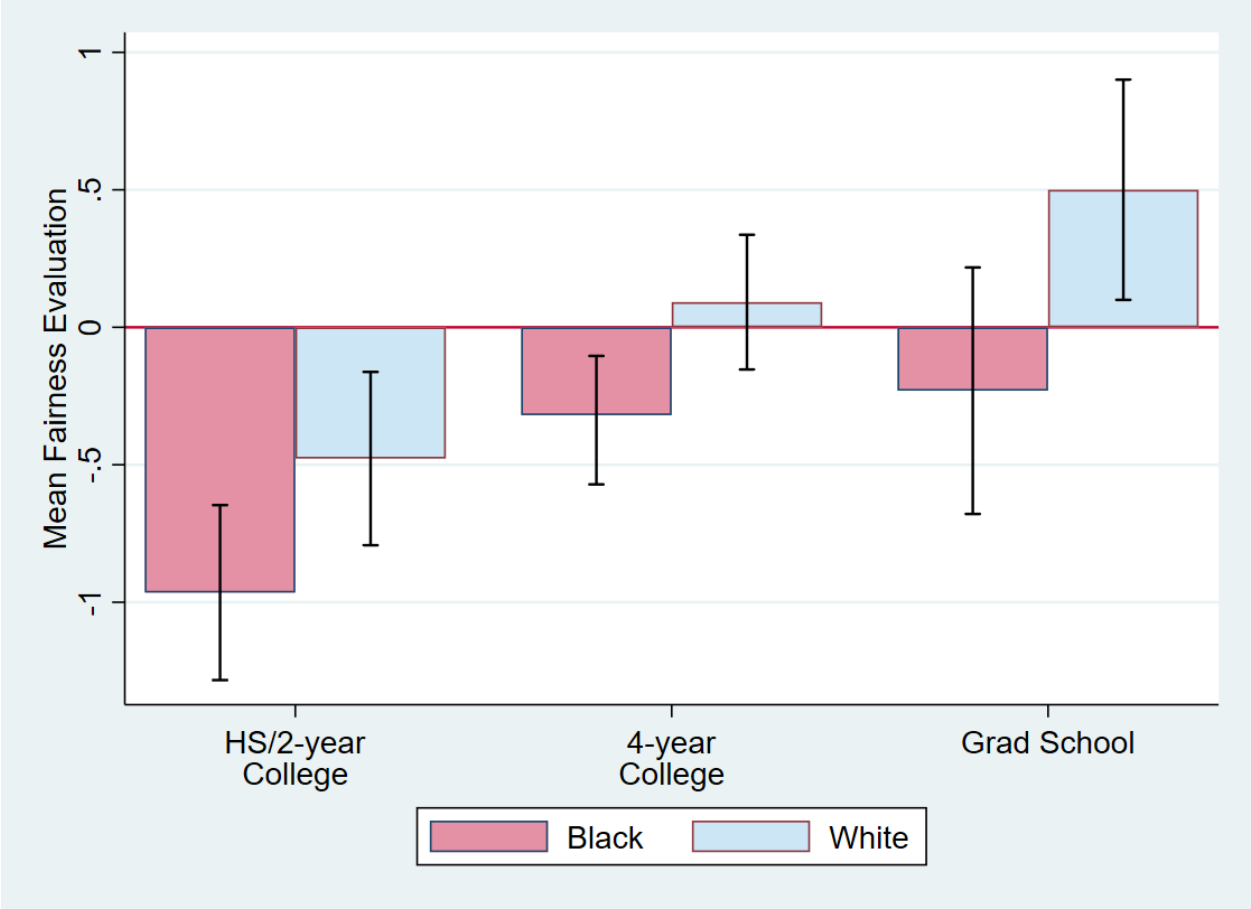


Notes: This figure is based on only Stage 1 observations. The *p*-values below are clustered by respondent.

- For conservative respondents:
  - HS/2-year vs. 4-year College = 0.304
  - 4-year College vs. Grad School = 0.199
  - Grad School vs. HS/2-year College = 0.506
- For moderate respondents:
  - HS/2-year vs. 4-year College = 0.032
  - 4-year College vs. Grad School = 0.564
  - Grad School vs. HS/2-year College = 0.457
- For liberal respondents:
  - HS/2-year vs. 4-year College = 0.001
  - 4-year College vs. Grad School = 0.974
  - Grad School vs. HS/2-year College = 0.008

Figure A4.2: Discriminatee Race Effects by Education

Despite being more tolerant of discriminatory acts in general, highly educated respondents react very similarly to the Race treatment.

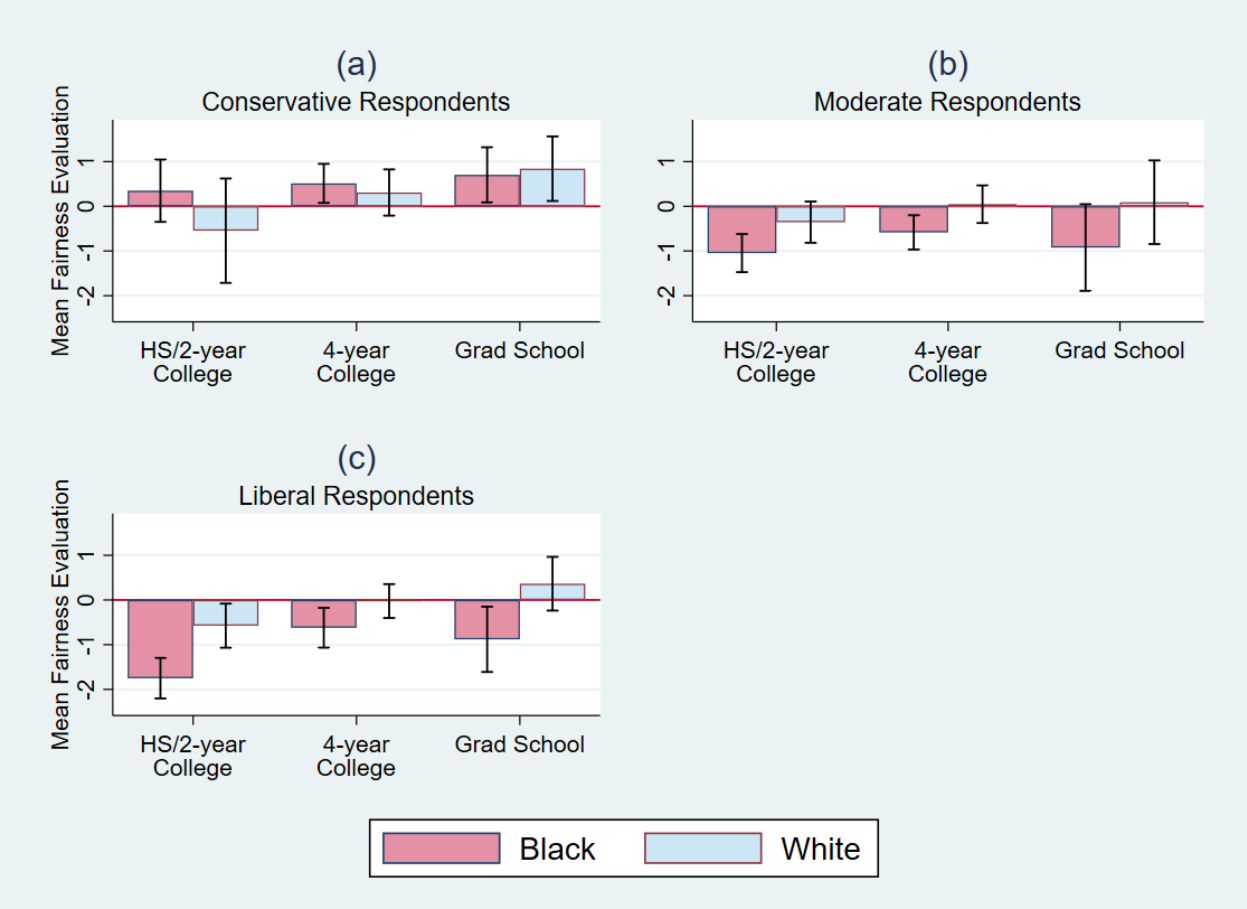


Notes: This figure is based on only Stage 1 observations. The *p*-values below are clustered by respondent.

- For HS/2-yar College graduates: Black vs. White = 0.032
- For 4-year College graduates: Black vs. White = 0.021
- For Graduate School graduates: Black vs. White = 0.016

Figure A4.3: Discriminatee Race Effects by Education and Political Leaning

The political difference in how respondents react to discriminatee race – moderates and liberals exhibit a discriminatee race effect and conservatives do not-- is present *within all three education groups.*



Notes: This figure is based on only Stage 1 observations. The *p*-values below are clustered by respondent.

- Conservative respondents:
  - For HS/2-yar College graduates: Black vs. White = 0.154
  - For 4-year College graduates: Black vs. White = 0.544
  - For Graduate School graduates: Black vs. White = 0.765
- Moderate respondents:
  - For HS/2-yar College graduates: Black vs. White = 0.029
  - For 4-year College graduates: Black vs. White = 0.028
  - For Graduate School graduates: Black vs. White = 0.107
- Liberal respondents:
  - For HS/2-yar College graduates: Black vs. White = 0.001
  - For 4-year College graduates: Black vs. White = 0.043
    For Graduate School graduates: Black vs. White = 0.009

Appendix 5: Robustness Tests for Sections 3 and 4

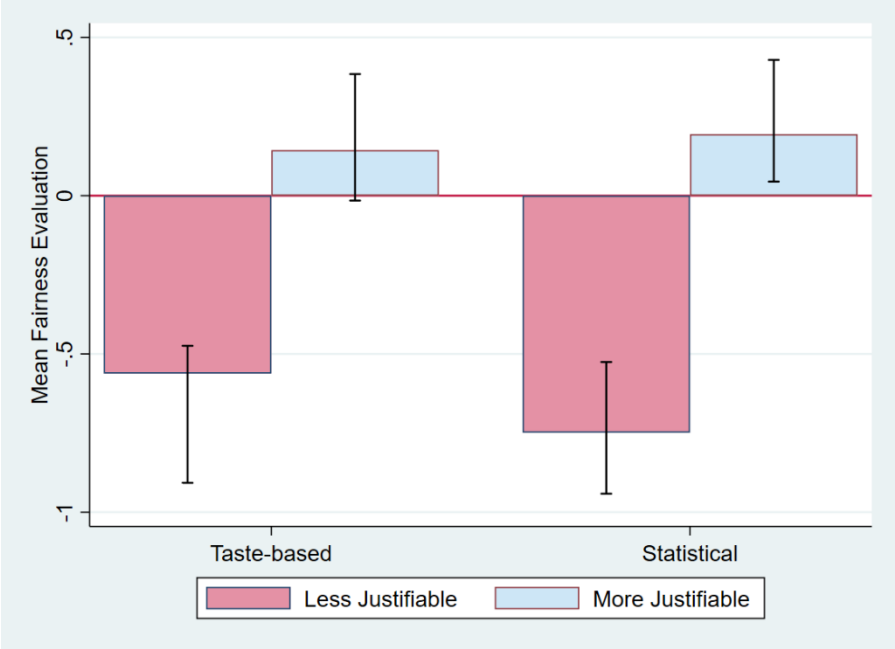
A5.1: Replicating Figures 2 and 3 Using First Scenarios Only

One of our more remarkable findings is that respondents' relative evaluations of the more versus less justifiable scenarios were so similar, regardless of the respondent's political orientation and of the race of the fictitious discriminatee. One might reasonably wonder whether this phenomenon reflects the fact that these two scenario types were always presented after each other and that subjects were asked to pay attention to the differences between the two types. To eliminate the possibility that subjects will be tempted to rank these two scenario types in the same way when they appear in sequence, we now replicate Figure 2 of the paper (which was estimated using both scenarios each person saw in Stage 1) using only data from the first scenario each respondent encountered. Remarkably, the results, shown in Figure A5.1.1, are indistinguishable from Figure 2. We conclude that subjects' perceptions of the relative fairness of the more- versus less-justifiable scenarios are the same, even when each subject has seen only one of the two scenario types.

Figure A5.1.2 repeats this same exercise for Figure 3, which illustrated discriminatee race effects using both scenarios each respondent encountered in Stage 1 of the survey. Figure A5.1.2 shows that the results are extremely similar if we use only information from the very first scenario each respondent encountered in the survey.

Figure A5.1.1: Fairness Ratings by Type of Discrimination and *Justifiability*: First Scenario Only

(replicates Figure 2)

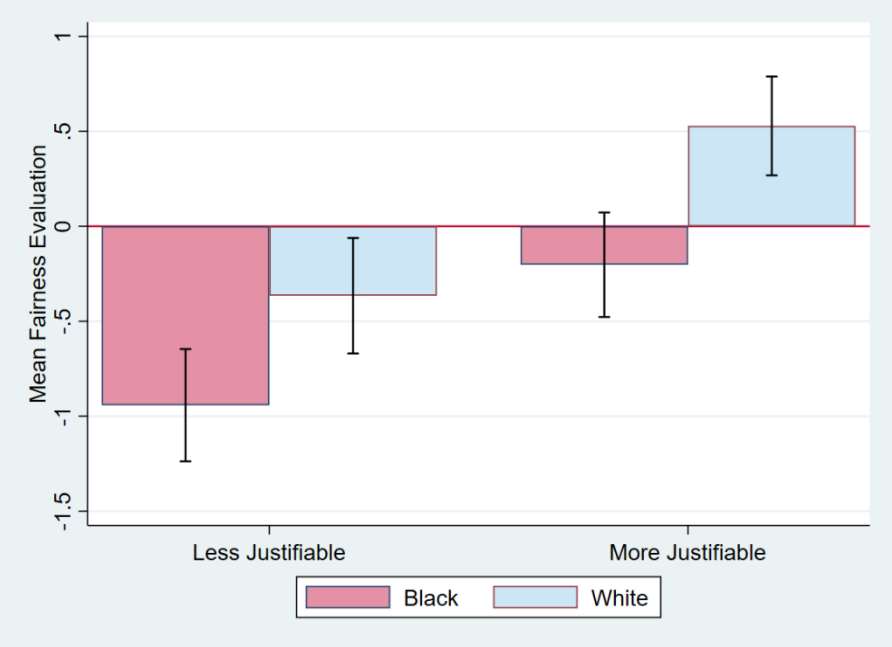


Notes: This figure replicates Figure 2 using only observations from the first scenario encountered by respondents in Stage One. Therefore, the *p*-values displayed below are not clustered.

- For taste-based discrimination, less vs. more justifiable scenarios = 0.001
- For statistical discrimination, less vs. more justifiable scenarios = 0.000
- Taste-based vs. statistical discrimination = 0.564

Figure A5.1.2: Fairness Ratings by *Justifiability* and Discriminatee Race: First Scenario Only

(replicates Figure 3)



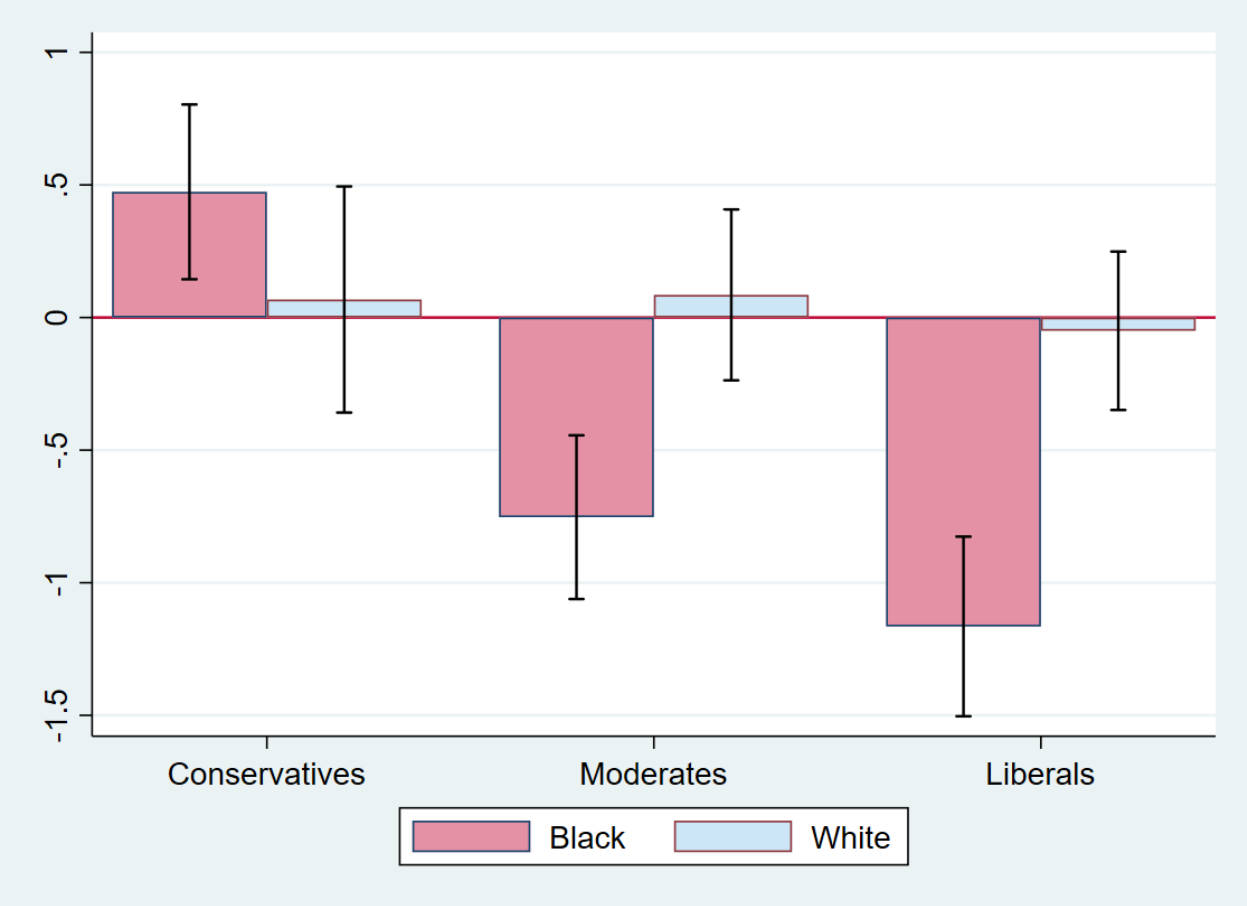
Notes: This figure replicates Figure 3 using only observations from the *first* scenario encountered by respondents in Stage One. Therefore, the *p*-values displayed below are not clustered.

- Black versus White Treatment
  - For less justifiable scenarios, Black versus White Treatment = 0.008
  - For more justifiable scenarios, Black versus White Treatment = 0.000
- More versus Less-*Justifiability* Treatment
  - For Black discriminatees, Less versus More-justifiable Treatment = 0.000 (difference = -0.7396)
  - For White discriminatees, Less versus More-justifiable Treatment = 0.000 (difference = -0.8943)
  - Less versus More *Justifiability* Gap equality across Black versus White treatment:
    - *p* = .5910

A5.2: Discriminatee Race Effects by Political Orientation for White Respondents Only

To probe the in-group bias hypothesis more deeply, here we replicate Figure 5 of the paper for White respondents only. The goal is to see if there is evidence of racial in-group bias if we focus on the subset of White respondents who label themselves as conservatives. Interestingly, the discriminatee race effect does switch signs in this group, relative to Figure 4 (which includes all respondents): conservative White respondents rate discrimination against Black people as *more* fair than discrimination against White people. This discriminatee race effect is not significantly different from zero at conventional levels, however (p=0.134).

Figure A5.2.1: Discriminatee Race Effects by Political Orientation, White Respondents Only



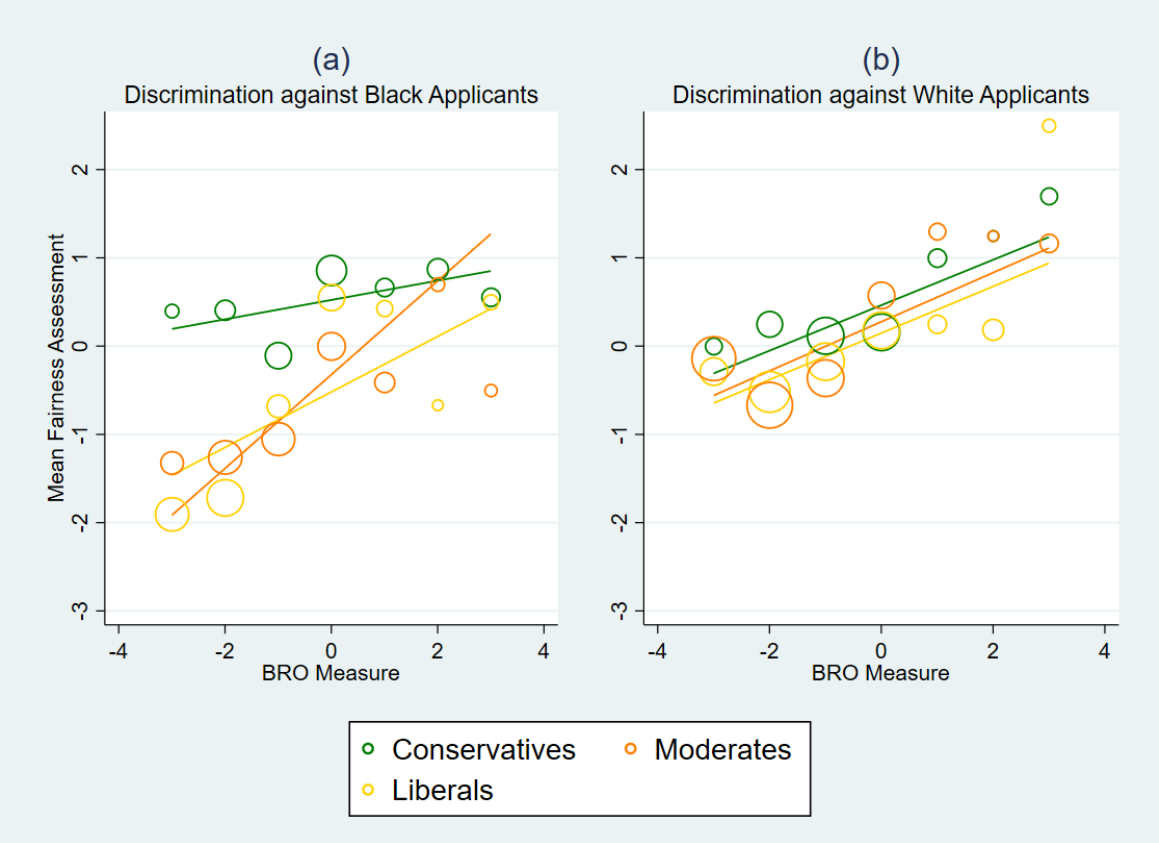
Notes: This figure reproduces Figure 6, but it only reflects the fairness evaluations of White respondents. The p-values below are clustered by respondent.

- For Conservatives, Black vs. White Treatment = 0.134
- For Moderates, Black vs. White Treatment = 0.000
- For Liberals, Black vs. White Treatment = 0.000

A5.3: Effects of Perceived Relative Opportunities (BRO) on Fairness Ratings, using Three Political Groups

Figure A5.3 replicates Figure 8 of the paper, showing separate results for moderates instead of combining moderates with liberals. For both anti-White and anti-Black discrimination moderates' fairness ratings are quite similar to liberals', and exhibit similar patterns with respect to BRO.

Figure A5.3: Effects of Perceived Relative Opportunities (BRO) on Fairness Ratings, by Discriminatee Race with Three Political Groups



Notes: This figure reproduces Figure 8, but it treats moderates and liberals as separate groups. Symbol size is proportional to the number of respondents. Sample is restricted to Stage 1 fairness assessments only. The *p*-values below are clustered by respondent, except for those pertaining to panel (c).

- Panel (a), Discrimination against Black Applicants
  - For Conservatives: slope = 0.109, $p = .218$
  - For Moderates, slope = 0.314, $p = .001$
  - Liberals, slope = 0.531, $p = .000$
- Panel (b), Discrimination against White Applicants
  - For Conservatives: slope = 0.257, $p = .094$
  - For Moderates, slope = 0.264, $p = .014$
  - Liberals, slope = 0.278, $p = .000$

Appendix 6: Analysis of Open-Text Responses

To gain some additional insights on respondent's motivations for their fairness assessments, we focused on two groups of respondents: those who indicated that the action in the last scenario they encountered was "unfair" or "very unfair" (211 respondents), and those who indicated that the action was "fair" or "very fair" (128 respondents). We then inspected all the open responses to this question:

> Recall the scenario that you just evaluated, in which [brief description of second scenario encountered in Stage 1].
>
> You thought that Michael's hiring decision was [very unfair / unfair / somewhat unfair / neither fair nor unfair / somewhat fair / fair / very fair]. In 50 words or less, please explain your response.

After eliminating respondents who entered "choose not to answer", responses that were undecipherable or consisted of unrelated text (presumably copied from the internet), and a small number of hard-to classify answers, this yielded 166 "unfair or very unfair" responses and 39 "fair or very fair" responses that could be assigned to three broad categories of reasons within each of these two groups.[2]

Tables 6.1 and 6.2 below summarize the counts of answers in each of these three categories, and provide examples of answers belonging to each category. Among the respondents who said discrimination was unfair or very unfair (Table 6.1), 51 percent (84/166) made a statement to the effect that making a hiring decision *based on race* was unfair. Another eight percent (14/166) said it was wrong to make a hiring decision on one's *tastes*. These reasons often overlapped (making it hard to choose which category was most appropriate). Both of them occurred much more often in the tasted-based scenarios. Finally, 41 percent (68/166) said that using statistical information was unfair (e.g. because each individual is different). Essentially all of these answers were for the statistical scenarios; many of them referred to the low quality of information in the less-justifiable statistical scenario. Words like racist, racism, bigoted, discrimination, prejudice, bias, and stereotype were commonly used in all these answers.

Among the respondents who said discrimination was fair or very fair (Table 6.2), missing and hard-to-interpret answers were much more common. With that caveat, 18 of 39 usable answers (46 percent) made a statement to the effect that a business owner's primary responsibility is to ensure their business thrives and survives. Almost all these answers referred to the customer discrimination

[2] Note that there were many more non-responses to the open-ended questions among respondents who thought discrimination was "fair" than "unfair". A spreadsheet containing all the open-ended responses submitted to the survey, indicating how we categorized the responses, and calculating all statistics presented in Appendix 6, can be downloaded at: https://docs.google.com/spreadsheets/d/1JsHVdvBWATU4MI88zLP-9RupOQsXIRnK/edit?usp=sharing&ouid=114674046533370433971&rtpof=true&sd=true

scenario, where catering to discriminatory customers allowed the employer to 'avoid losing sales). Another 36 percent (14/39) referred to an employer's rights (for example, to hire whomever he wishes, regardless of the reason). Finally, 18 percent (7/39) said that that it was acceptable to make hiring decisions based on statistical information on productivity. All of these responses referred to statistical discrimination scenarios. Notably, however, almost half of them referred to the low-justifiability version, where the hiring decision was based on hearsay. For example, "Well, it was based on some sort of evidence-based reasoning process rather than just a sentiment of not wanting to work with a White person."

Table A6.1 Summary of stated reasons why discrimination was "unfair" or "very unfair"

	Count of responses		
Reason:	**Taste-Based Scenarios**	**Statistical Scenarios**	**All Scenarios**
Wrong to use race	67	17	84
Wrong to use information	8	60	68
Wrong to use tastes	13	1	14
Total	88	78	166

Note: 166 responses that fit these three categories were obtained from 211 respondents selecting "unfair" or "very unfair" on the last scenario they encountered. 15 of the remaining answers were "prefer not to answer"; the rest could not be easily classified, including undecipherable text and irrelevant text copied from the web. 62 of the responses contained at least one word from the following list: racist, racism, bigoted, discrimination, prejudice, bias, or stereotype.

Examples of "wrong to use race" statements:

"He should hire black people anyways regardless of his feeling because it is the right thing to do. Regardless of how people feel about interacting with black people, the employer has an obligation to be fair in hiring practices."

"I think it's unfair that you decide against hiring someone just because you don't like interacting with people of that race."

"Someone's ability to be hired should never be based off of the color of their skin or opinions of others."

Note: a large majority of these statements occurred in the taste-based treatments.

Examples of "wrong to use information" statements:

"He was going off of information that was basically gossip with his neighbor."

"I feel like because he is basing who to hire on information and statistics about local black workers, which he got from other owners. I don't see that as fair because everyone is different."

"It's crazy that a professional person would make a hire based on what a neighbor said. It's really racial profiling and not at all based on worker skills or experience."

Note: a large majority of these statements occurred in the statistical treatments.

Examples of "wrong to use tastes" statements:

"It is insane not to hire an employee simply because you do not like people of their race. The individual shouldn't be judged based on racist views."

"Their preferences are racist and should not be taken into consideration. Those customers need to overcome their racist tendencies, it is not the responsibility of the business to cater to them."

“I think it's unfair to avoid hiring an individual because you didn't enjoy interacting with other individuals from their race.”

Note: a large majority of these statements occurred in the taste-based treatments.

Table A6.2. Summary of stated reasons why discrimination was "fair" or "very fair"

Reason:	**Taste-Based Scenarios**	**Statistical Scenarios**	**All Scenarios**
Business must thrive	17	1	18
Statements about employer rights	8	6	14
OK to raise profits using statistical information	0	7	7
Total:	25	14	39

Note: 39 responses that fit these three categories were obtained from 128 respondents selecting "unfair" or "very unfair" on the last scenario they encountered. 36 of the remaining answers were "prefer not to answer"; the rest could not be deciphered, were irrelevant text (presumably copied from the web), or not easily classifiable.

Examples of "business must thrive" statements:

"The hiring decision was fair because any individual in Michael's shoes would do anything within their power to protect their business by all means necessary."

"If clients do not like to interract (sic) with white personnel that means that white workers hurt business."

"He needs to retain his customers, so he should listen to what they want to see in employees, even if their responses are a little uncomfortable."

Note: almost all of these statements (16/17) were for the customer discrimination scenario (more-justifiable, taste-based)

Examples of "employer rights" statements:

"It's his company he can hire whoever he choses (sic). He does not have to give an answer to anyone or share his hiring views. He can choose what is best at any time without answering to anyone."

"Andrew does run the business so it is within his rights to not hire a black man because he doesn't enjoy interacting with them."

"The employer should have the right to hire who he is most comfortable with regardless of the reasons."

Examples of "OK to use statistical information":

"Michael's hiring decision was fair because he collected details about Black workers and their problems and decided to choose white employer (sic)."

"Data and reliable statistical proof is respected in every other type of research and information gathering, why wouldn't it carry weight in this type of situation as well?"

“Well, it was based on some sort of evidence-based reasoning process rather than just a sentiment of not wanting to work with a White person.”

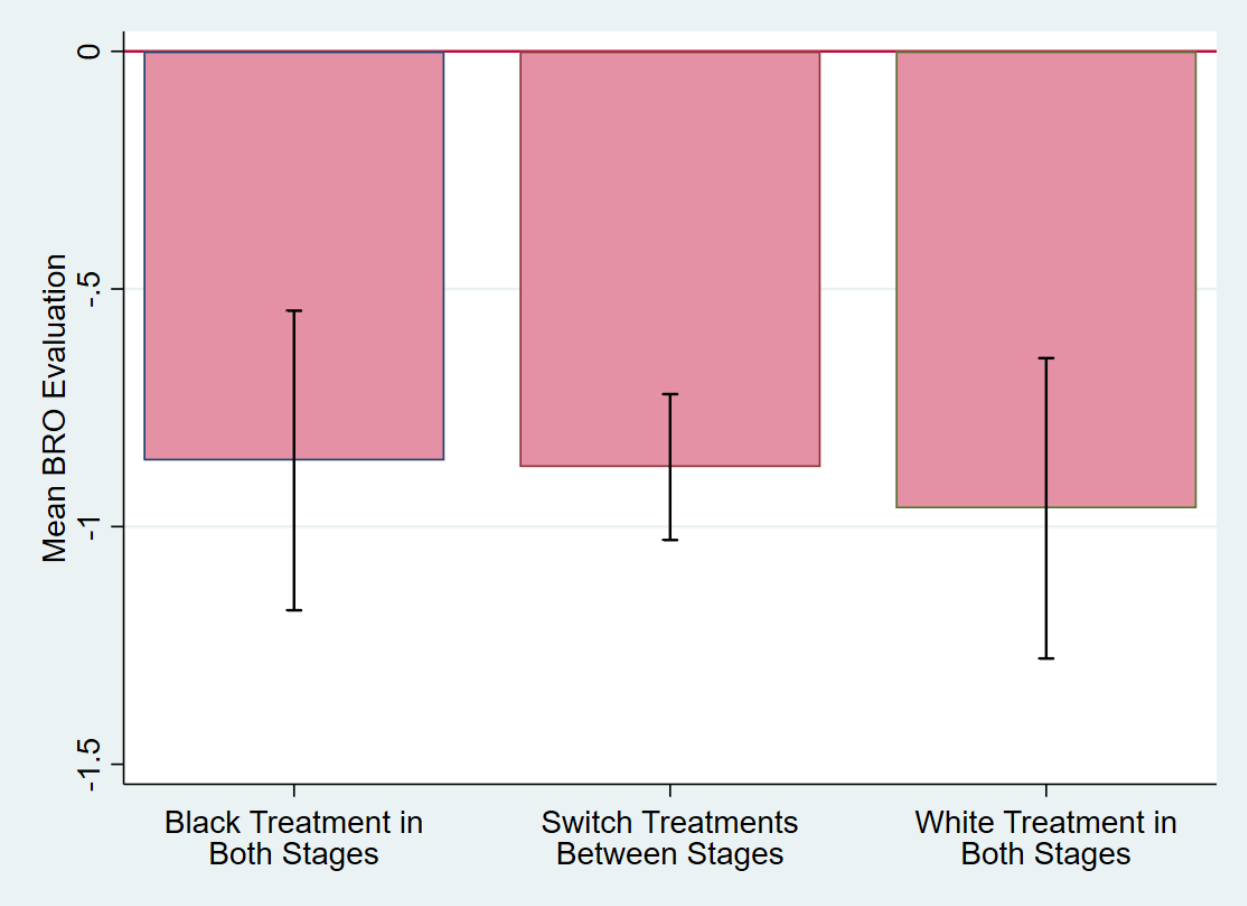
Note: All of these statements (7/7) were for statistical discrimination scenarios.

Appendix 7: Experimenter Demand Effects do not Explain the Race Treatment Order Effects

In Section 5.1 of the paper, we proposed an explanation of the observed order effects for the Black Treatment based on experimenter demand effects. According to this hypothesis, subjects who first encounter a Black (White) discriminatee assume the experimenters are liberals (conservatives), and then provide fairness assessments they think will please liberals (conservatives). In this Appendix we test this hypothesis by arguing that subjects who want to please the experimenters should also tailor their answers to other survey questions to please the experimenters. In this regard, the survey questions that seem most likely to be susceptible to such manipulation are (a) subjects' assessments of Black peoples' relative economic opportunities (BRO), and (b) subjects' reported political orientations. This Appendix demonstrates that subjects' answers to these questions are not influenced by which discriminatee races they encountered earlier in the survey, suggesting that experimenter demand effects probably do not account for the order effects we see in subjects' fairness assessments.

Specifically, Figure A7.1 reports the mean assessment of Black peoples' relative economic opportunities (BRO) for three groups of respondents: respondents who encountered the Black treatment in both Stages, respondents who encountered the White treatment in both Stages, and subjects who encountered a mix of Black and White treatments. The differences between the three groups are all small and statistically insignificant. Figure A7.2 replicates the analysis for subjects' reported political leaning (on a scale from -3 to +3). Finally, Figure A7.3 repeats this analysis separately for the share of subjects reporting a Democratic or Republican party preference. In all cases, the effects of being previous exposure to White versus Black experimental treatments are small and statistically insignificant.

Figure A7.1: Mean BRO Evaluation Across Respondents' Survey Experience



Notes*:*

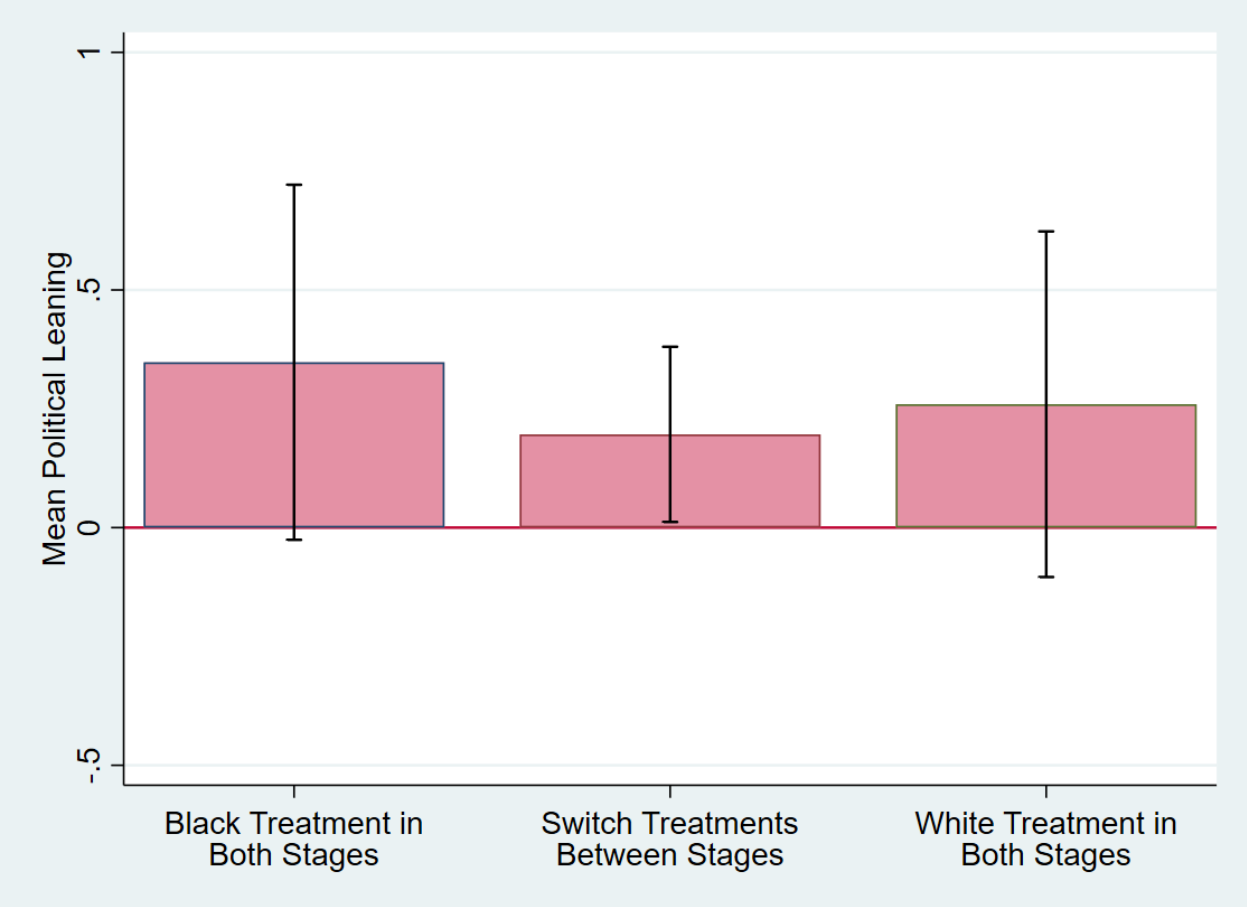
BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more).

The *p*-values below are clustered by respondent.

- Black Treatment in Both Stages vs Switchers = 0.938
- Switchers vs White Treatment in Both Stages = 0.624
- Black Treatment in Both Stages vs White Treatment in Both Stages = 0.655

If the respondents choose their BRO reports to cater to the (inferred) political preferences of the experimenters, we should see a monotonic increase in BRO from left to right. Such an increase is not present.

Figure A7.2: Mean Political Leaning Across Respondents' Survey Experience



Notes*:*

Political leaning is the respondent's self-description on a scale of -3 (very conservative) to 3 (very liberal).

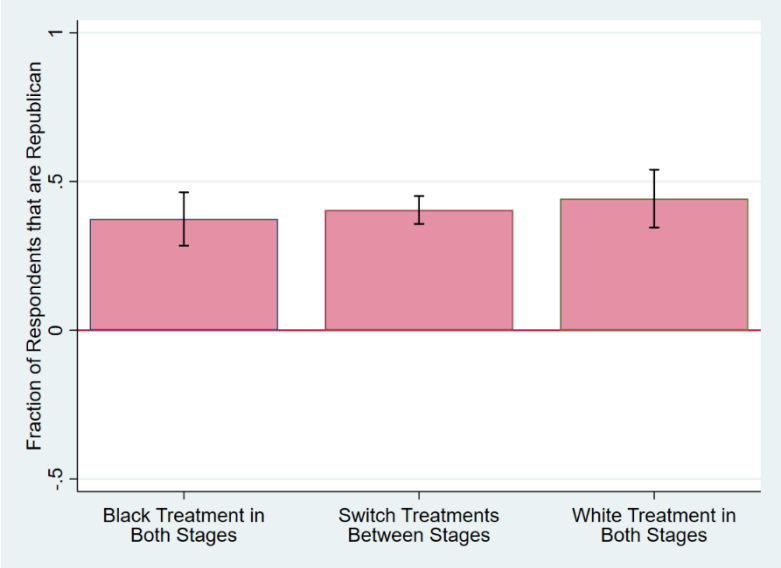
Notes*:* The *p*-values below are clustered by respondent.

- Black Treatment in Both Stages vs Switchers = 0.471
- Switchers versus White Treatment in Both Stages = 0.758
- Black Treatment in Both Stages vs White treatment in Both Stages = 0.737

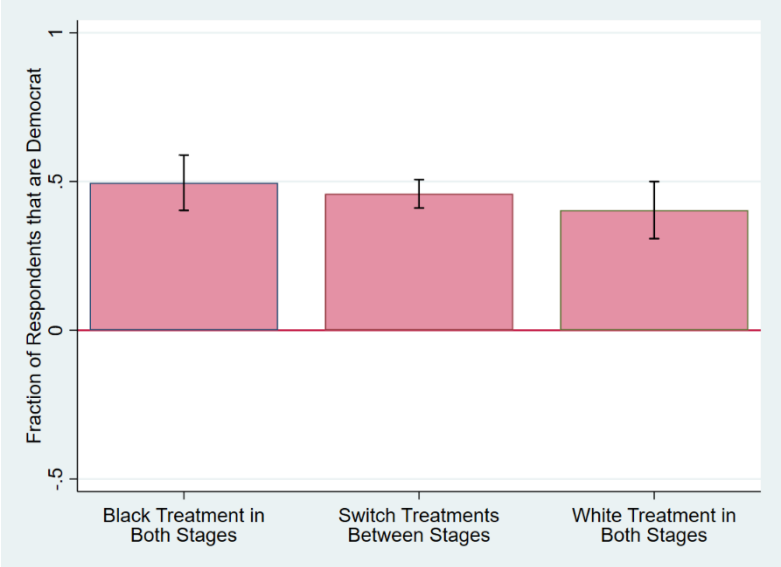
If the respondents modify their reported political leanings to cater to the (inferred) political preferences of the experimenters, we should see a monotonic decrease (shift from liberal towards conservative) from left to right. Such a decrease is not present.

Figure A7.3 Reported Party Preference Across Respondents' Survey Experience

(a)



(b)



Notes*:* The *p*-values below are clustered by respondent.

- For the fraction of Republican respondents:
  - Black Treatment in Both Stages vs Switchers = 0.553
  - Switchers versus White Treatment in Both Stages = 0.484
  - Black Treatment in Both Stages vs White treatment in Both Stages = 0.305
- For the fraction of Democrat respondents:
  - Black Treatment in Both Stages vs Switchers = 0.482
  - Switchers versus White Treatment in Both Stages = 0.310
  - Black Treatment in Both Stages vs White treatment in Both Stages = 0.173

Appendix 8: Estimating α

A8.1 Splitting the Sample by Groups 1 and 2 (*Business Rights Advocates* versus *Utilitarians*)

In this Section, we first document how the *race* treatment order effect differs between respondent Groups 1 and 2. We show that these order effects are absent in Group 1 (*the Business Rights Advocates*). In Group 2 (the *Utilitarians*) the order effects are even stronger than in the aggregate data. We next provide data that allow us to operationalize the 'trade-off' model of Group 2's ratings changes in Section 5.3 of the paper. Specifically, Figures A8.1.1 and Figures A8.1.2 replicate Figure A3.4.1 (which showed that subjects' Stage 2 fairness assessments depend on the *race* treatment they encountered in Stage 1) separately for Groups 1 and 2.

Figure A8.1.1 shows the Stage 2 mean fairness ratings of respondents in Group 1, disaggregated by the *race* treatments they encountered in both Stages of the experiment. Perhaps the most noteworthy feature is that all the fairness assessments are positive (discrimination is more fair than unfair), but small in value: All the means are between 0 (neither fair nor unfair) and 1 (somewhat fair). Closely related, Group 1's fairness assessments do not respond to the *race* treatments, nor do they depend on the order in which the treatments are administered. Specifically, we cannot reject that Group 1's Stage 2 fairness assessments are unaffected by the treatment they encountered in Stage 1 (p = .582 for the Black treatment in Stage 2; p = .769 for the White treatment in Stage 2).

Turning to Group 2, Figure A8.1.2 shows a very different pattern. Now all the fairness assessments are negative, but their magnitude is strongly related to the race of the discriminatee. Figure A8.1.2 also shows that Group 2's Stage 2 fairness ratings *do* depend on the discriminatee race they encountered in Stage 1. Specifically, respondents who encountered the Black treatment in Stage 2 rated it as much less fair if they also encountered it in Stage 1 than if they encountered a White discriminatee in Stage 1; this difference is highly statistically significant (p=.013).

Finally, we apply a simple fairness reporting model to the preceding data to estimate the relative weight Group 2 assigns to their utilitarian preferences, compared to race-blindness. The model's key identifying assumption is that respondents are not aware of their desires to be race-blind until they encounter a race treatment switch in the experiment. We estimate that members of Group 2 place roughly equal weight on these two fairness criteria.

Figure A8.1.1: *Race* Treatment Order Effects for Group 1 (*Business Rights Advocates*: all conservatives, plus moderates and liberals with BRO ≥ 0)

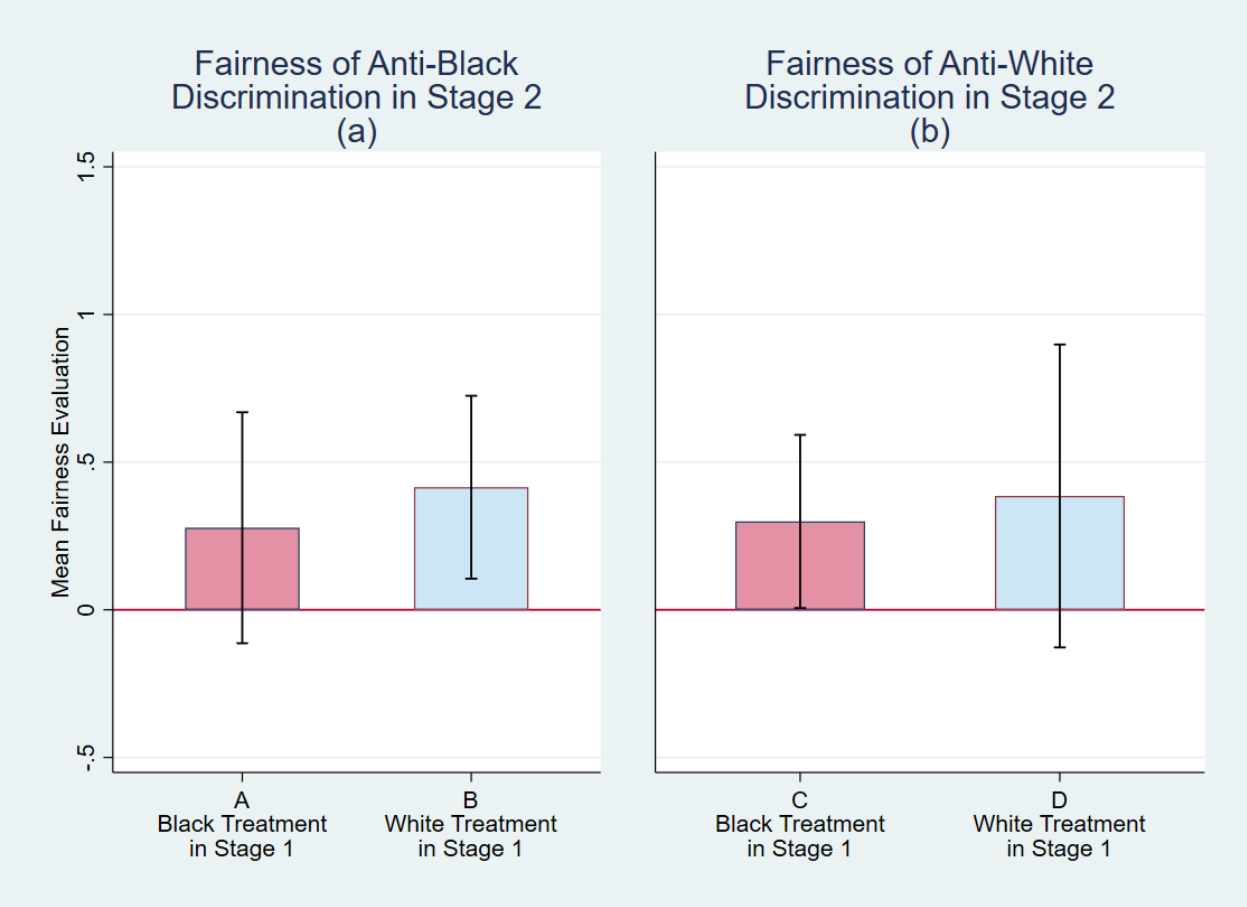


Figure A8.1.1 shows that Group 1's Stage 2 fairness ratings *do not* depend on the discriminatee race they encountered in Stage 1:

***p*-values**:

A vs B = 0.582

C vs D = 0.769

A vs C = 0.930

B vs D = 0.921

Notes: All *p*-values are clustered by respondent.

Figure A8.1.2: *Race* Treatment Order Effects for Group 2 (*Utilitarians*: moderates and liberals with BRO<0)

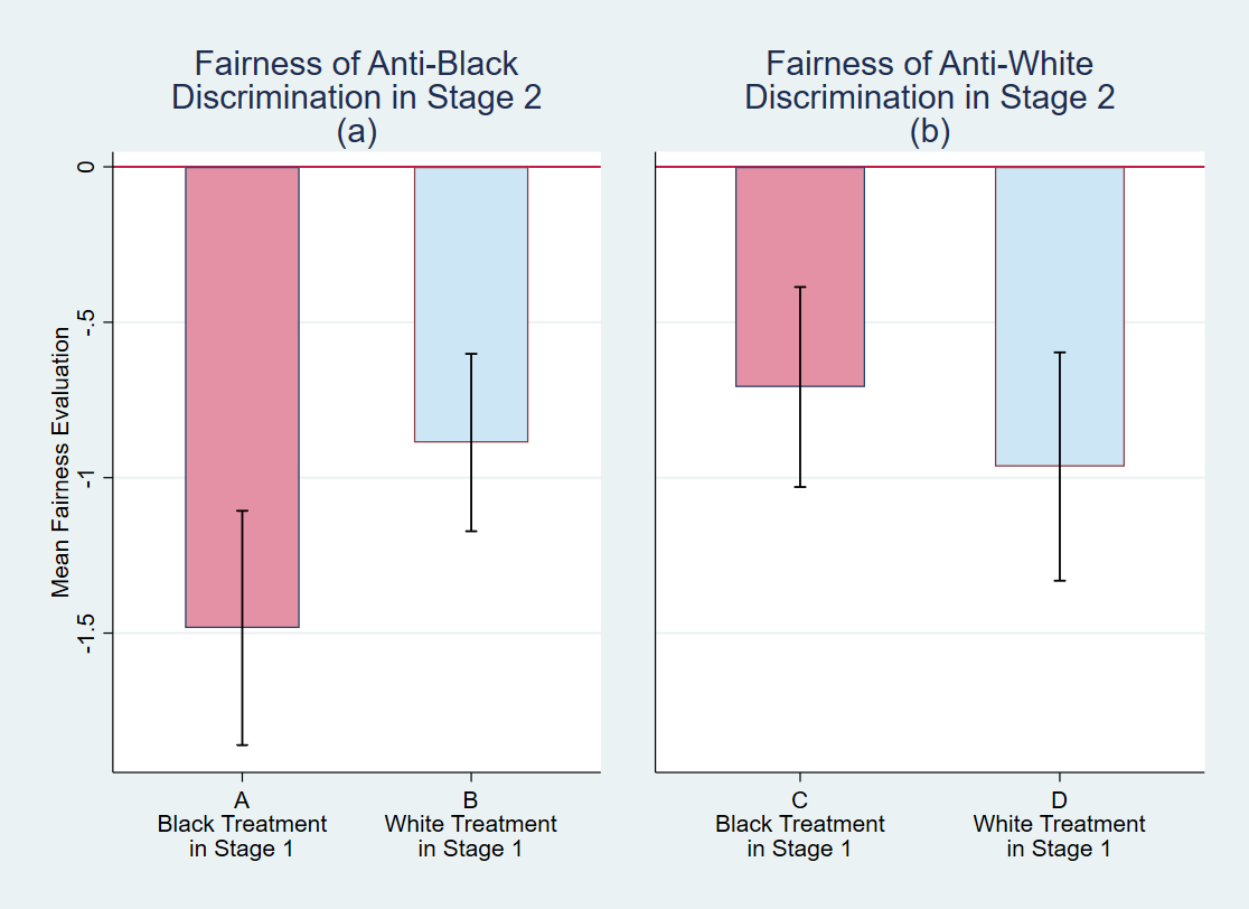


Figure A8.1.2 shows that Group 2's Stage 2 fairness ratings *do* depend on the discriminatee race they encountered in Stage 1. Specifically, respondents who encountered the Black treatment in Stage 2 rated it as much less fair if they encountered a Black discriminatee in Stage 1 than if they encountered a White discriminatee in Stage 1:

***p*-values**:

A vs B = 0.013

C vs D = 0.269

A vs C = 0.002

B vs D = 0.740

Notes: All *p*-values are clustered by respondent.

To calculate the relative weight assigned by Group 2 to their 'true' utilitarian fairness rating, we assume that subjects' Stage 1 assessments, B_i^1 and W_i^1represent their "true" utilitarian ratings in a setting where they don't need to consider race-blindness (B_i^* and W_i^*). In Stage 2, race treatment *switchers* then face a conflict. For example, White-to-Black switchers could either:

- Report their true rating of discrimination against the *new* group ($B_i^2 = B_i^*$).
- Report the same rating they assigned in Stage 1 ($B_i^2 = W_i^1$).

If switchers assign a weight α to their true rating, the Stage 2 ratings of W-to-B switchers will be:

$$B_i^2 = \alpha B_i^* + (1-\alpha)W_i^1 \qquad (1)$$

where:

- B_i^* is their individual, true assessment of anti-Black discrimination (not observed).
- W_i^1 is their assessment of anti-White discrimination in Stage 1 (observed).

While $B_i^*$is not observed for W-to-B switchers, for any pre-defined group (e.g. Group 2), random treatment assignment allows us to estimate its sample mean ($\bar{B}^*$) from subjects who received the Black treatment in Stage 1. Using this 'trick', we can calculate α (separately) for W-to-B switchers and B-to-W switchers, yielding:

$\alpha = 0.49$ for the White-to-Black switchers. (roughly equal weight)

$\alpha = 0.68$ for the Black-to-White switchers more weight on the 'truth')

Statistically:

- For W-to-B switchers, we can reject both α=0 and α=1. (*p*=.000, *p*=.004)
- For B-to-W switchers, we reject both α=0 but not α=1. (*p*=.000, *p*=.098)
- We cannot reject $\alpha = 0.5$ for either type of switcher (*p*=.969, *p*= .220).

Thus, members of Group 2 behave as if they place about equal weight on utilitarian and race-blind fairness criteria. Confidence intervals for α can be calculated separately for W-B switchers and B-W switchers as:

W-to-B Switchers: [0.243,0.800]

B-to-W Switchers: [0.405, 1.075]

Thus, among W-to- B switchers (where the order effect is strongest) we can reject both $\alpha = 0$ and $\alpha = 1$.

A8.2 Splitting the Sample by Political Leaning (conservatives versus [moderates + liberals])

In this Section, we replicate Appendix 8.1, splitting the sample by self-reported political affiliation instead of Groups 1 versus 2 (as defined in Section 4.4). Since Groups 1 and 2 are predominantly conservative and moderate/liberal respectively, all the results are very similar. Like Group 2, moderates and liberals exhibit a highly significant *Race* treatment order effect (which we would expect since all Group 2 members are moderate or liberal) and conservatives exhibit no such effect (which we expect since Group 1 is mostly conservative). The estimates of α for [moderates + liberals] are very similar to those for Group 2 as well.

Figure A8.2.1: *Race* Treatment Order Effects for Conservative Respondents

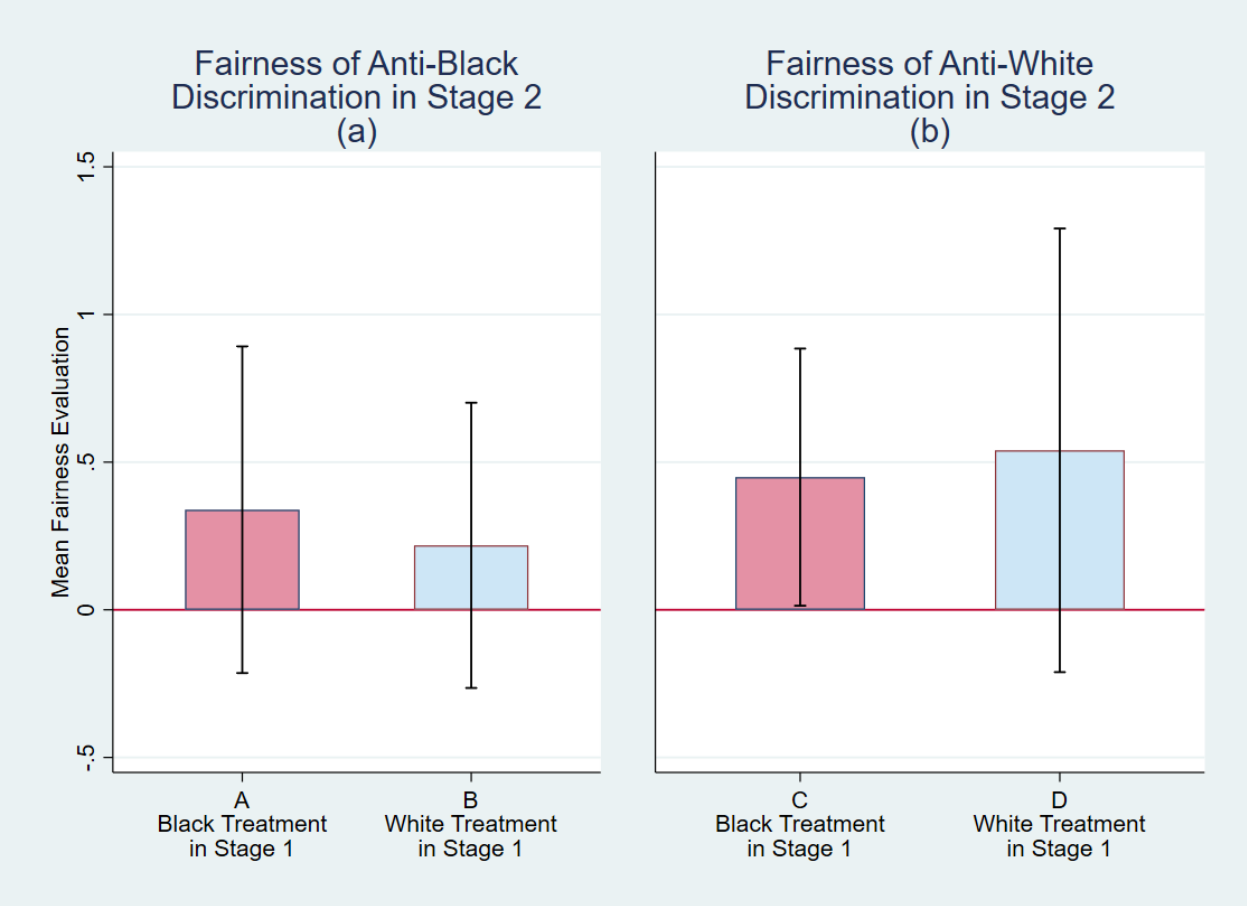


Figure A8.2.1 shows that conservative respondents' Stage 2 fairness ratings *do not* depend on the discriminatee race they encountered in Stage 1:

***p*-values**:

A vs B = 0.739

C vs D = 0.829

A vs C = 0.750

B vs D = 0.460

Notes: All *p*-values are clustered by respondent.

Figure A8.2.2: *Race* Treatment Order Effects for Moderate and Liberal Respondents

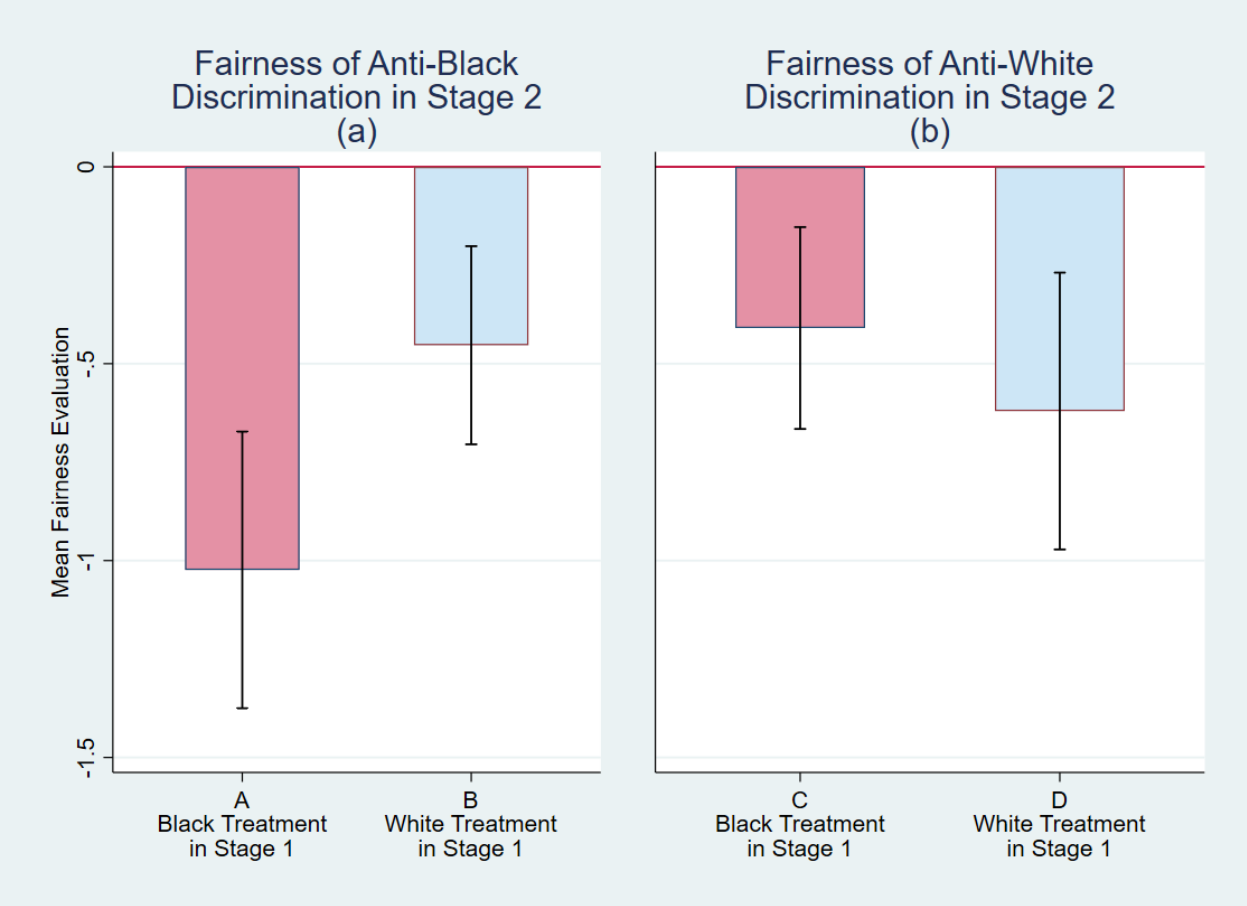


Figure A8.2.2 shows that moderate and liberal respondents' Stage 2 fairness ratings *do* depend on the discriminatee race they encountered in Stage 1. Specifically, respondents who encountered the Black treatment in Stage 2 rated it as much less fair if they also encountered it in Stage 1 than if they encountered a White discriminatee in Stage 1:

***p*-values**:

A vs B = 0.009

C vs D = 0.336

A vs C = 0.005

B vs D = 0.443

Notes: All *p*-values are clustered by respondent.

Using the same method as in Appendix A8.1, we can again calculate α (separately) for W-to-B switchers and B-to-W switchers (among moderate and liberals respondents), yielding:

$\alpha = 0.44$ for the White-to-Black switchers. (slightly more weight on RBRs)

$\alpha = 0.62$ for the Black-to-White switchers (slightly more weight on the 'truth')

Statistically:

- For W-to-B switchers, we can reject both α=0 and α=1. (*p*=.003, *p*=.007)
- For B-to-W switchers, we reject both α=0 but not α=1. (*p*=.000, *p*=.067)
- We cannot reject $\alpha = 0.5$ for either type of switcher (p=.678, p= .423).

Thus, moderates and liberals behave as if they place about equal weight on utilitarian and race-blind fairness criteria**.** Confidence intervals for α can be calculated separately for W-B switchers and B-W switchers as:

W-to-B Switchers: [0.155,0.791]

B-to-W Switchers: [0.348, 1.033]

Thus, among W-to-B switchers (where the order effect is strongest) we can reject both $\alpha = 0$ and $\alpha = 1$.

Appendix 9: Replicating the Main Figures with ACS Weights

In this Appendix, we replicate Figures 2-8 with a set of post-stratification weights. These weights were derived from the 2019 American Community Survey (ACS). They re-weight our MTurk responses by the relative prevalence of our respondents in the ACS in 24 cells, defined by gender (male and female), race (White versus non-White), education (HS/2-year college versus 4-year college or higher) and age (18-24 versus 25-44 versus 45 years of age or older). Table A9.1 shows the share of respondents in our MTurk sample (unweighted), in our weighted MTurk sample, and in the ACS. We do not re-weight the sample on political leaning here because the ACS does not contain that information.

Columns 1 and 3 of Table A9.1 show the sample composition of our MTurk respondents and 2019 ACS respondents at least 18 years old. They show that men and White respondents are modestly over-represented on MTurk. People between the ages of 25 and 44 and four-year college graduates are highly over-represented. Column 2 shows that our weights do quite a good job of correcting for these forms of non-representativeness.

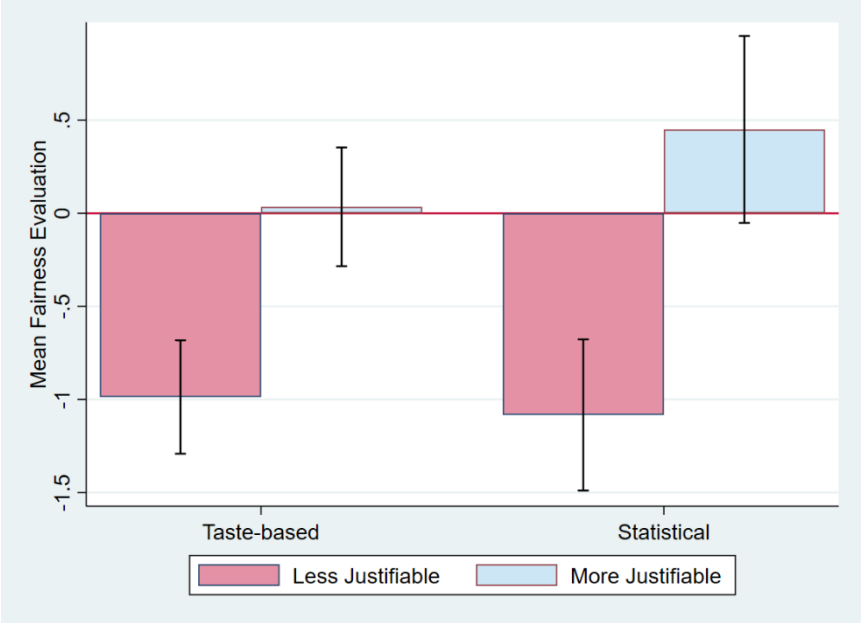
The remaining exhibits in this Appendix replicate Figures 2-8 using these weights. All the main patterns discussed in the paper are also present here, with one small exception: the weak positive association between BRO and the fairness of discrimination among conservative respondents in Figure 8(a) becomes somewhat stronger and statistically significant. Similar to Figure 8, however, the slope for conservatives remains much lower than the slope for moderates / liberals.

Table A9.1: Raw and Re-Weighted Sample composition, ACS weights.

CHARACTERISTIC	MTurk Sample (1)	Weighted Sample (2)	2019 ACS Sample (3)
Male	0.600	0.522	0.487
Female	0.400	0.478	0.513
White respondents	0.780	0.673	0.628
Non-White respondents	0.115	0.327	0.372
Age 18-24	0.037	0.128	0.119
Age 25-44	0.729	0.368	0.343
Age 45 and over	0.234	0.504	0.538
HS or less, or 2-year/some college	0.294	0.671	0.694
4-year college or graduate school	0.706	0.329	0.307
Observations	642	642	2,599,171

Notes: Column 1 contains the percentage of respondents across various demographic characteristics within the MTurk sample. Column 2 contains these percentages for the 2019 American Community Survey (ACS) sample for comparison.

Figure A9.1: Fairness Ratings by Type of Discrimination and *Justifiability* (replicates Figure 2)



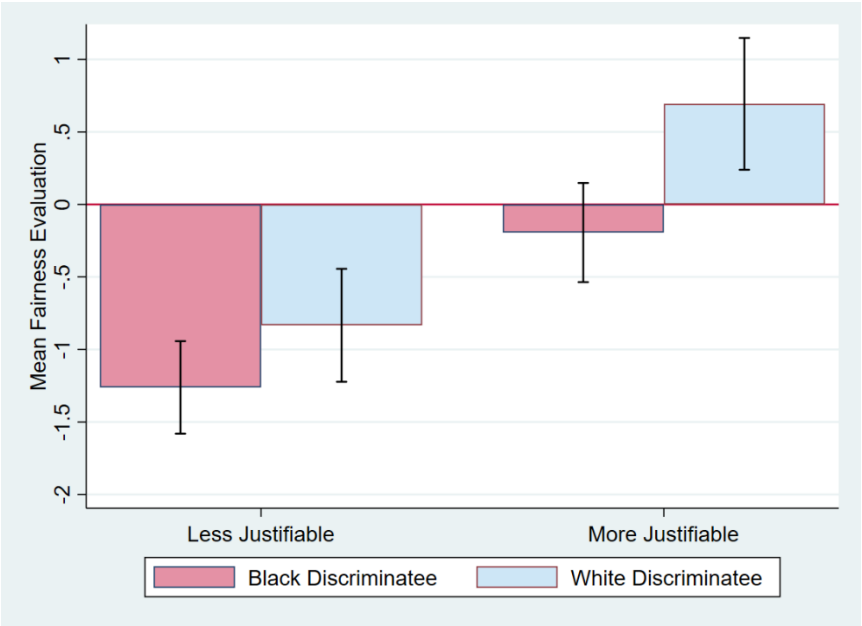
p-values:

Less- versus more justifiable treatments:
Overall: *p*=.000
Within taste-based: *p*=.000
Within statistical: *p*=.000

Taste versus Statistical Discrimination:
Overall: *p*=.505
Within Less-Justifiable: *p*=.709
Within More-Justifiable: *p*=.170

Note: This figure is based on only Stage 1 observations. All p-values are clustered by respondent.

Figure A9.2: Fairness by *Justifiability* and Discriminatee Race (replicates Figure 3)



p-values:

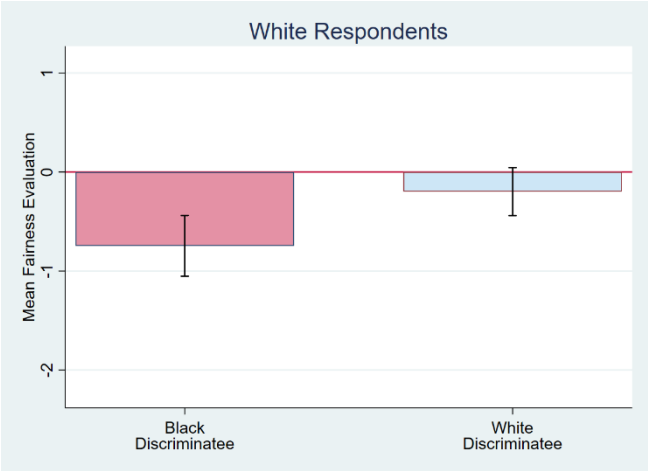
Black versus White Treatment:
Overall: *p*=.003
Within Less-Justifiable: *p*=.095
Within More-Justifiable: *p*=.002

Less versus More Justifiable Treatment:
Overall: *p*=.000
Within Black Discriminatees: *p*=.000
Within White Discriminatees: *p*=.000

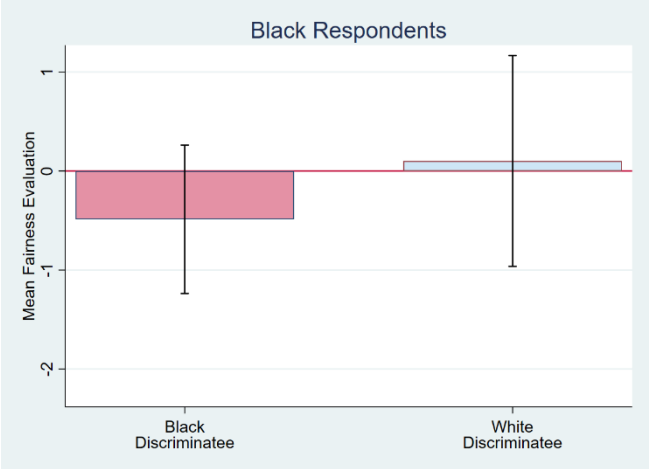
Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 1.068 units less fair. Within White Discriminatees, less-justifiable scenarios are 1.527 units less fair. A test for equality of the Less versus More *Justifiability* Gap between the Black and White treatment yields *p* = .140.

Figure A9.3: Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 4)

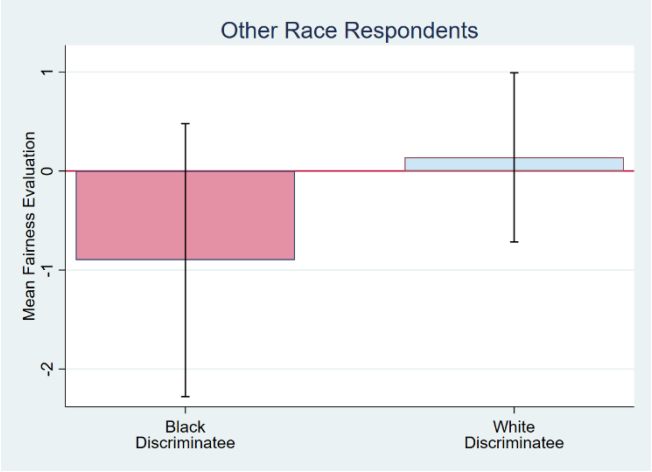
(a)



(b)



(c)

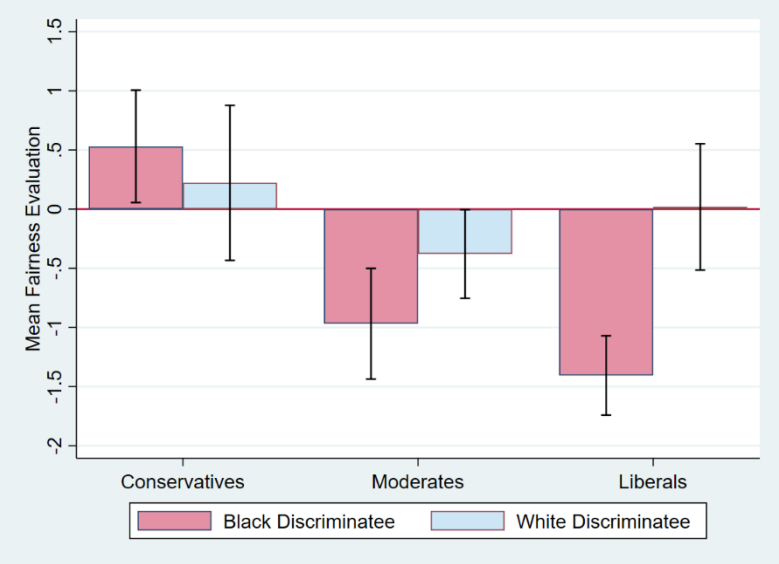


p-values:

Black versus White Treatment:
Overall: Overall: *p*=.003
Within White respondents: *p*=.006
Within Black respondents: *p*=.360
Within Other respondents: *p*=.195

Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) across all three racial groups yields *p* = .832

Figure A9.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Figure 5)



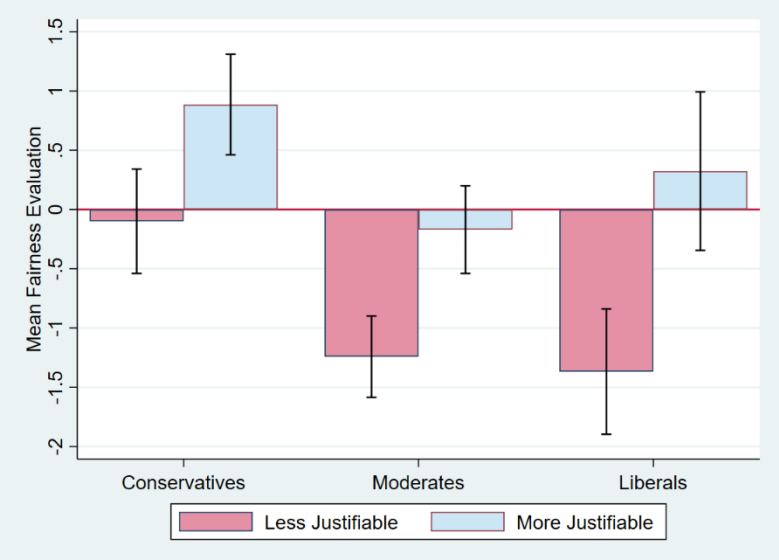
p-values:

Black versus White Treatment:

Overall: *p*=.003
Within Conservatives: *p*=.449
Within Moderates: *p*=.052
Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields *p* = .058. A test for equality between conservatives and (moderates + liberals) yields *p* = .006.

Figure A9.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent's Political Leaning (replicates Figure 6)



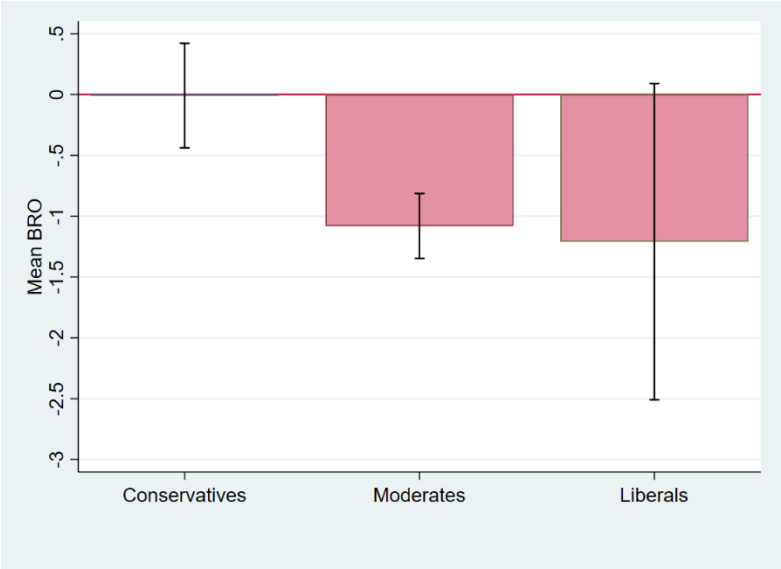
p-values:

Less versus More Justifiable Treatment:

Overall: *p*=.000
Within Conservatives: *p*=.000
Within Moderates: *p*=.000
Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the Less versus More *Justifiability* Gap across Conservatives, Moderates, and Liberals yields *p* = .153.

Figure A9.6: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 7)



p-values:

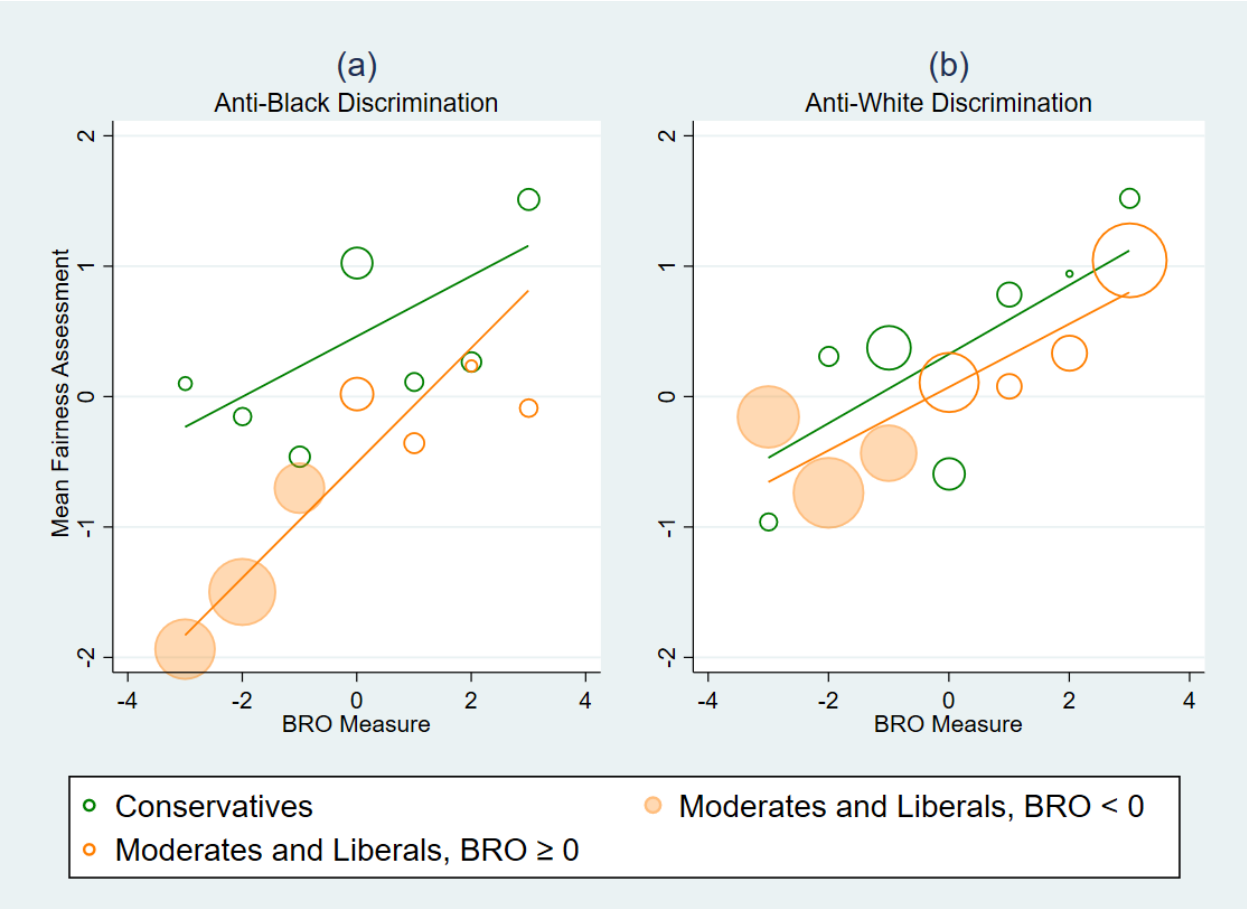
Conservatives vs. Moderates: *p*=.000

Moderates vs. Liberals: *p*=.848
Liberals vs. Conservatives *p*=.084

Notes:

BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more). This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of BRO across all three political groups yields $p = .010$.

Figure A9.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 8)



Notes: Symbol size is proportional to the number of respondents. Sample is restricted to Stage 1 fairness assessments only. The *p*-values below are clustered by respondent, except for those pertaining to Panel (c).

- Panel (a), Discrimination against Black Applicants
  - For Conservatives: slope = 0.232, $p = .021$
  - For Moderates and Liberals, slope = 0.441, $p = .000$
- Panel (b), Discrimination against White Applicants
  - For Conservatives: slope = 0.265, $p = .204$
  - For Moderates and Liberals, slope = 0.242, $p = .000$

Appendix 10: Replicating the Main Figures with GSS Weights

In this Appendix, we replicate Figures 2-8 with an alternative set of post-stratification weights. These weights were derived from the 2020 General Social Survey (GSS), and they are based only on a 7-point political leaning scale (i.e., extremely conservative, conservative, slightly conservative, moderate, slightly liberal, liberal, and extremely liberal). Columns 1 and 3 of Table A10.1 show the sample composition of our MTurk respondents and 2020 GSS respondents at least 18 years old. Overall, MTurk respondents differ from the GSS in two main ways: First, compared to the GSS a smaller share of MTurk respondents choose the middle three categories: 'moderate' or 'slightly' liberal / conservative, while MTurkers are also more likely to locate in the two 'extreme' categories. In this sense, MTurkers are politically more extreme than GSS respondents.[3] Second, almost identical shares of MTurkers and GSS respondents choose some degree of conservative leaning (ranging from slight to extreme), but many more MTurkers choose some liberal leaning (47.3 versus 30.2 percent). Thus, on average, MTurkers are also more liberal than the U.S. population as a whole. Because our weights do not interact political leaning with any other characteristics, the weighted MTurk sample in column 2 of Table A10.1 mimics the GSS sample perfectly.[4]

The remaining exhibits in this Appendix replicate Figures 2-8 using these weights. All the main patterns discussed in the paper are also present here. The one exception noted with the ACS weights in Appendix 9 does not occur here, suggesting that the unusual political mix of MTurkers is not responsible for *any* of the main results in the paper.

[3] It is possible, however, that some of this is caused by a difference in phrasing of the middle category between the two surveys. See Appendix 2 for additional details.

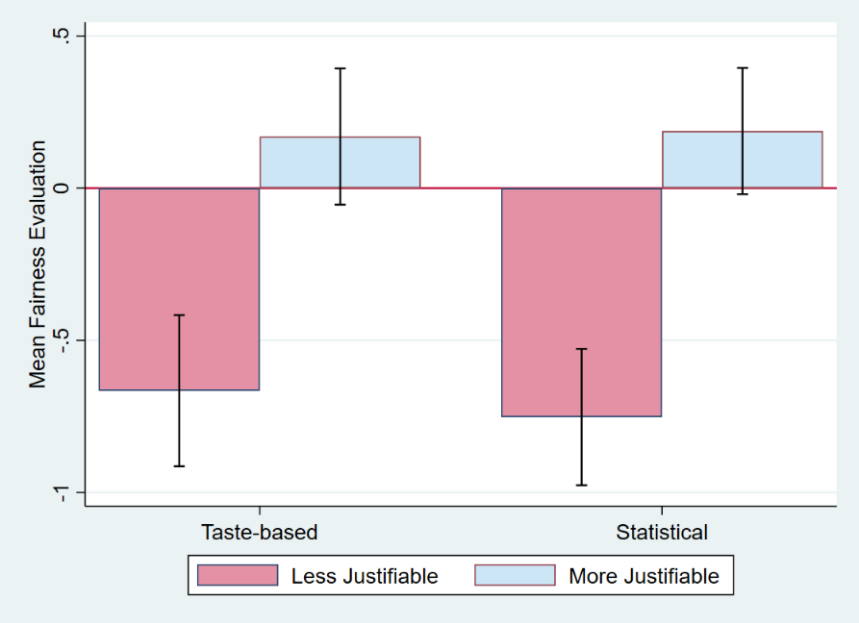
[4] Because of the small size of the MTurk and GSS samples, we did not re-weight our MTurk sample to mimic GSS demographic characteristics; attempts to do this yielded highly extreme and imprecise weights.

Table A10.1: Raw and Re-Weighted Sample composition, GSS weights.

CHARACTERISTIC	MTurk Sample (1)	Weighted Sample (2)	2020 GSS Sample (3)
Extremely conservative	0.101	0.051	0.051
Conservative	0.164	0.168	0.168
Slightly conservative	0.092	0.146	0.146
Moderate	0.170	0.332	0.332
Slightly liberal	0.095	0.121	0.121
Liberal	0.274	0.132	0.132
Extremely liberal	0.104	0.049	0.049
Observations	642	642	1,776

Notes: Column 1 contains the percentage of respondents across various demographic characteristics within the MTurk sample. Column 2 contains these percentages for the 2019 American Community Survey (ACS) sample for comparison.

Figure A10.1: Fairness Ratings by Type of Discrimination and *Justifiability* (replicates Figure 2)



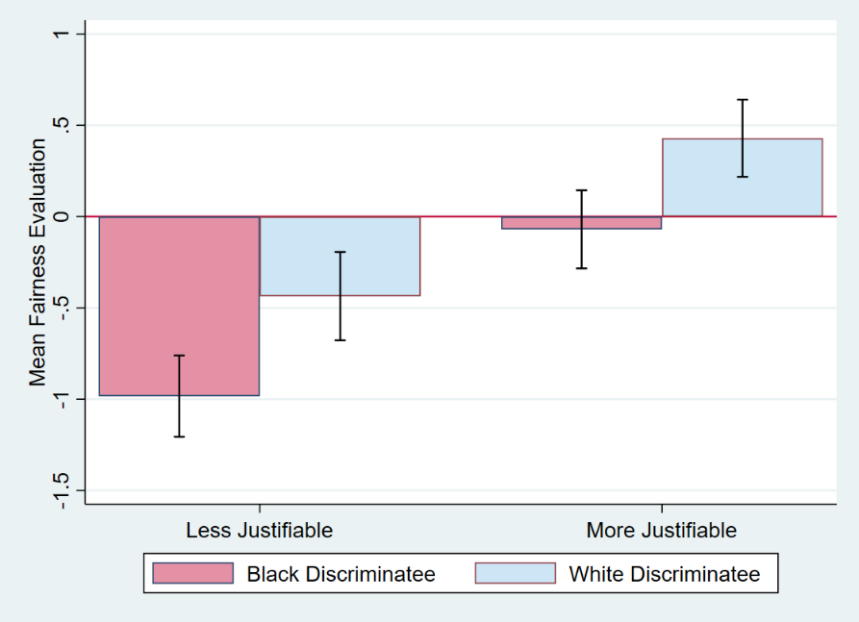
p-values:

Less- versus more justifiable treatments:
Overall: *p*=.000
Within taste-based: *p*=.000
Within statistical: *p*=.000

Taste versus Statistical Discrimination:
Overall: *p*=.813
Within Less-Justifiable: *p*=.610
Within More-Justifiable: *p*=.907

Note: This figure is based on only Stage 1 observations. All p-values are clustered by respondent.

Figure A10.2: Fairness by *Justifiability* and Discriminatee Race (replicates Figure 3)



p-values:

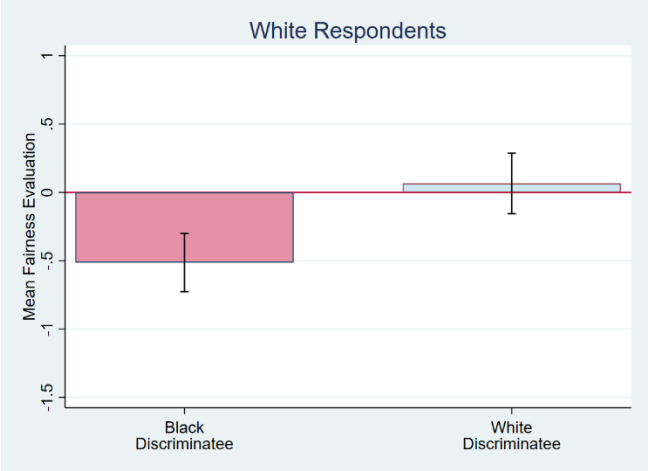
Black versus White Treatment:
Overall: *p*=.000
Within Less-Justifiable: *p*=.001
Within More-Justifiable: *p*=.001

Less versus More Justifiable Treatment:
Overall: *p*=.000
Within Black Discriminatees: *p*=.000
Within White Discriminatees: *p*=.000

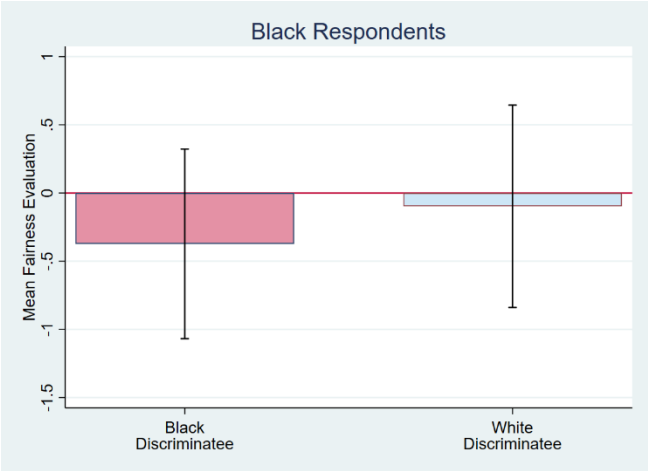
Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.914 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.865 units less fair. A test for equality of the Less versus More *Justifiability* Gap between the Black and White treatment yields *p* = .744.

Figure A10.3: Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 4)

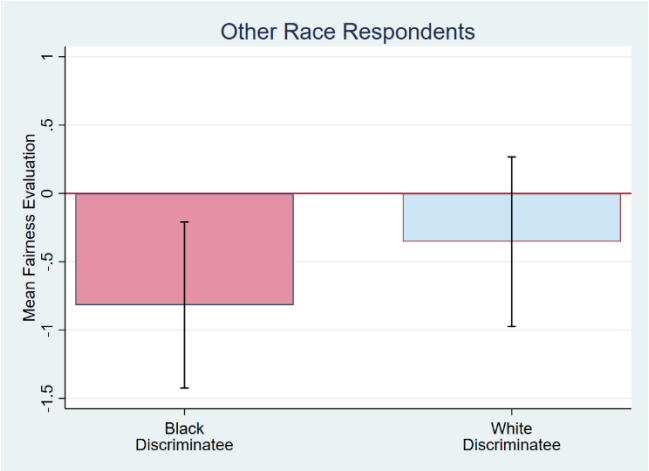
(a)



(b)



(c)



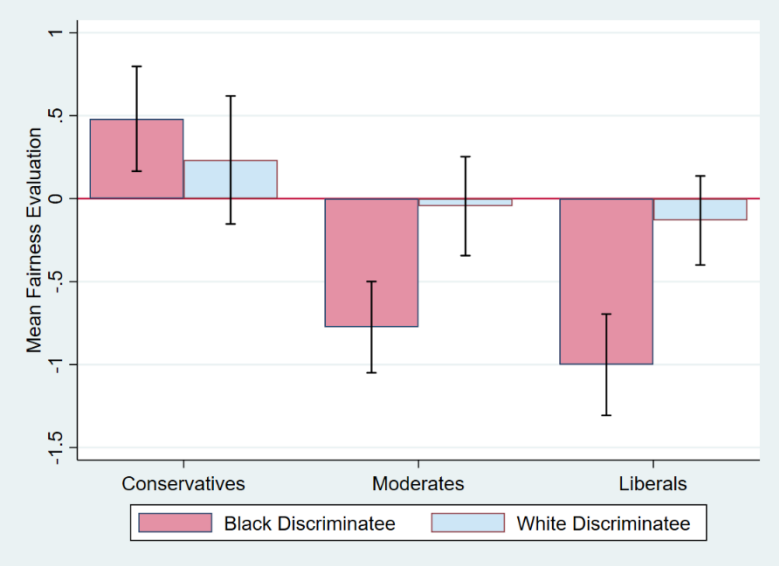
p-values:

Black versus White Treatment:

Overall: Overall: *p*=.000
Within White respondents: *p*=.000
Within Black respondents: *p*=.583
Within Other respondents: *p*=.279

Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e. the Black treatment) across all three racial groups yields *p* = .827.

Figure A10.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Figure 5)

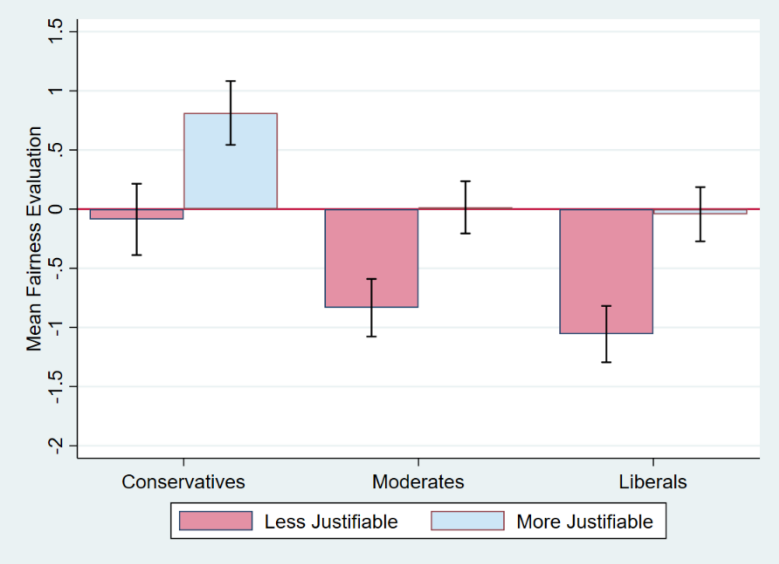


p-values:

Black versus White Treatment:
Overall: *p*=.000
Within Conservatives: *p*=.323
Within Moderates: *p*=.000
Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields *p* = .628. A test for equality between conservatives and (moderates + liberals) yields *p* = .001.

Figure A10.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent's Political Leaning (replicates Figure 6)

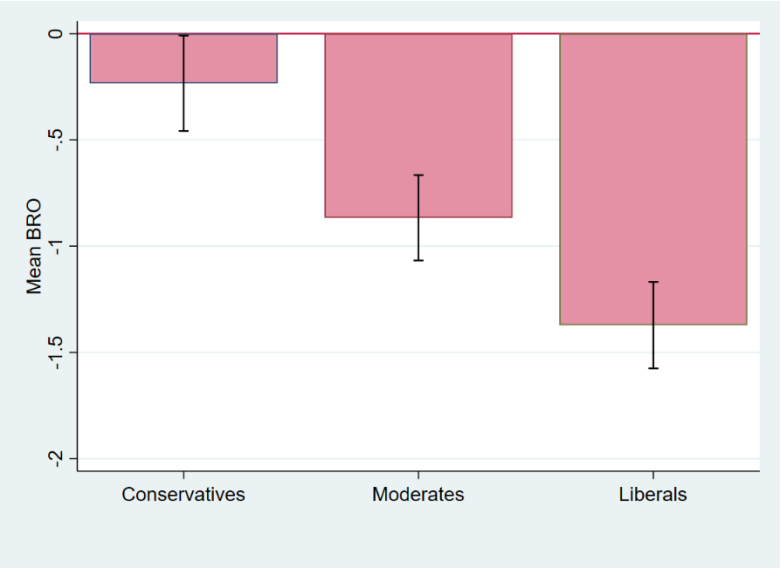


p-values:

Less versus More Justifiable Treatment:
Overall: *p*=.000
Within Conservatives: *p*=.000
Within Moderates: *p*=.000
Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the Less versus More *Justifiability* Gap across Conservatives, Moderates, and Liberals yields *p* = .541.

Figure A10.6: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 7)



p-values:

Conservatives vs. Moderates: *p*=.000

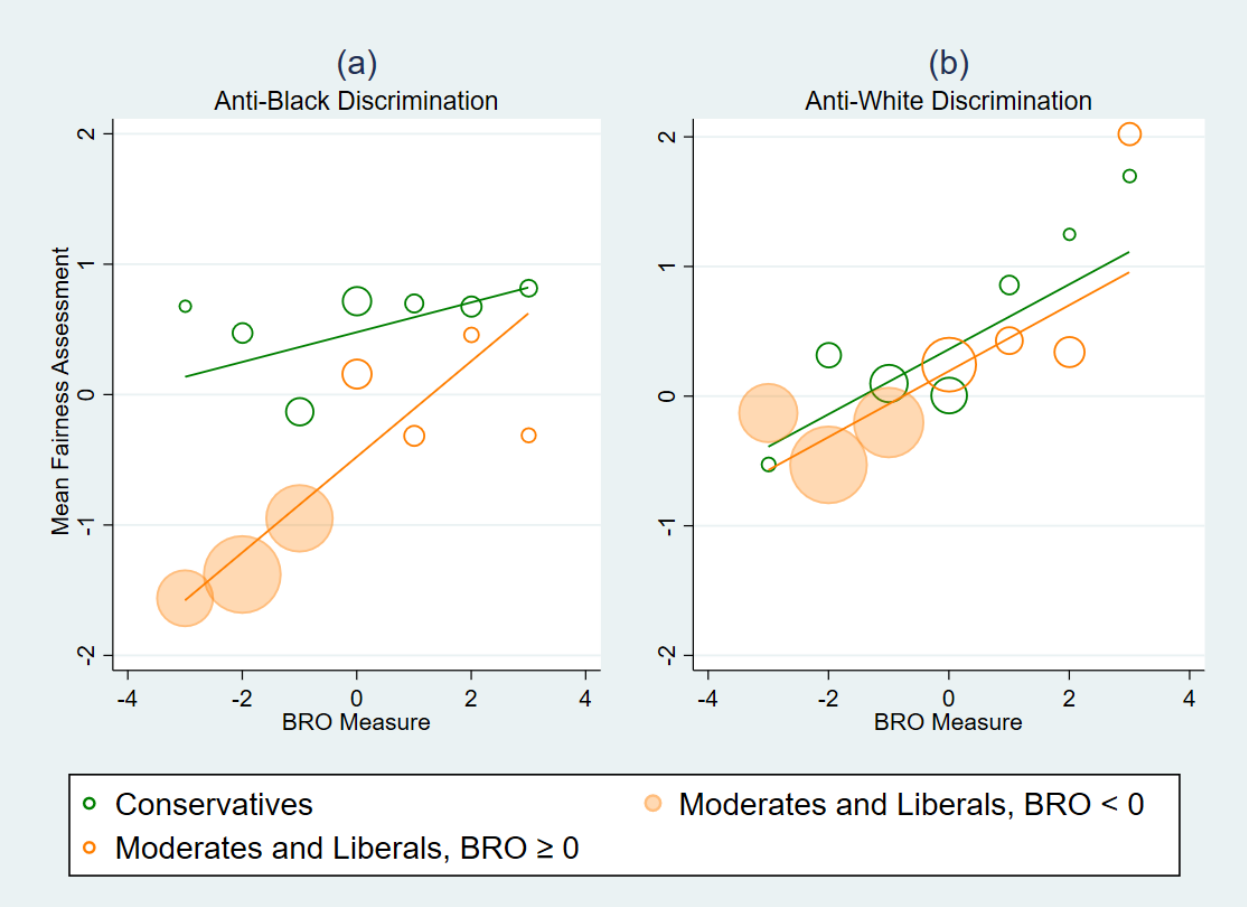
Moderates vs. Liberals: *p*=.001

Liberals vs. Conservatives *p*=.000

Notes:

BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more). This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of BRO across all three political groups yields $p = .505$.

Figure A10.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 8)



Notes: Symbol size is proportional to the number of respondents. Sample is restricted to Stage 1 fairness assessments only. The *p*-values below are clustered by respondent, except for those pertaining to Panel (c).

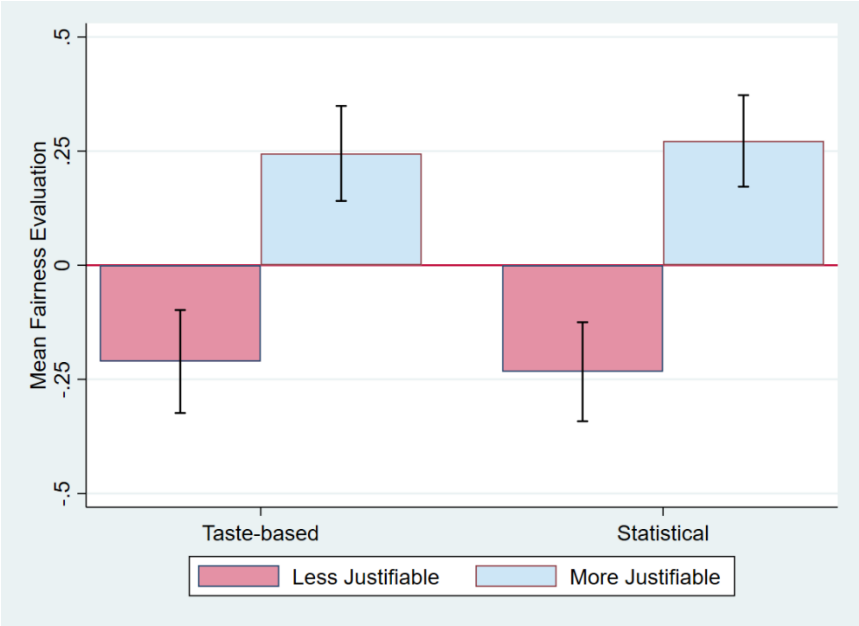
- Panel (a), Discrimination against Black Applicants
  - For Conservatives: slope = 0.114, p = .231
  - For Moderates and Liberals, slope = 0.367, p = .000
- Panel (b), Discrimination against White Applicants
  - For Conservatives: slope = 0.250, p = .096
  - For Moderates and Liberals, slope = 0.254, p = .001

Appendix 11: Replicating the Main Figures with Standardized Fairness Measures

In this section, we replicate the main figures and table by using a standardized version of our fairness ratings. Therefore, all of the means displayed in Figures 2-8 illustrate deviations from the mean fairness rating for the entire sample, i.e., -0.286 on a scale of -3 to 3 where the standard deviation is 1.920.[5] We also standardize the BRO (Black relative opportunity) measure, where its mean is -0.886, also on a scale of -3 to 3 where the standard deviation is 1.498. In short, all of the figures are comparable to the ones using the raw fairness and BRO measures.

[5] Specifically, we standardize our fairness evaluation measures with respect to the full sample.

Figure A11.1: Fairness Ratings by Type of Discrimination and *Justifiability* (replicates Figure 2)



p-values:

Less- versus more justifiable treatments:

Overall: *p*=.000

Within taste-based: *p*=.000

Within statistical: *p*=.000

Taste versus Statistical Discrimination:

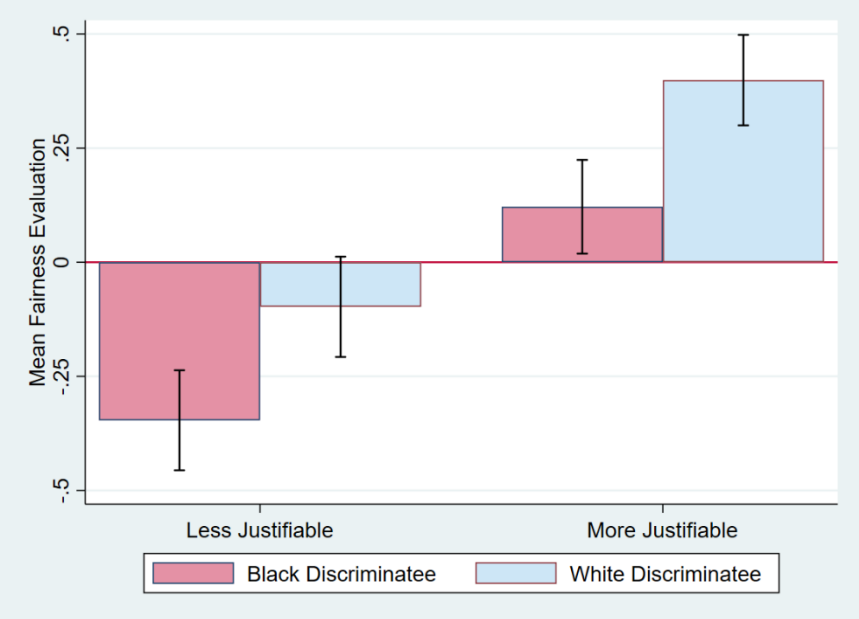
Overall: *p*=.971

Within Less-Justifiable: *p*=.779

Within More-Justifiable: *p*=.710

Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent.

Figure A11.2: Fairness by *Justifiability* and Discriminatee Race (replicates Table 3)



p-values:

Black versus White Treatment:

Overall: *p*=.000

Within Less-Justifiable: *p*=.002

Within More-Justifiable: *p*=.000

Less versus More Justifiable Treatment:

Overall: *p*= .000

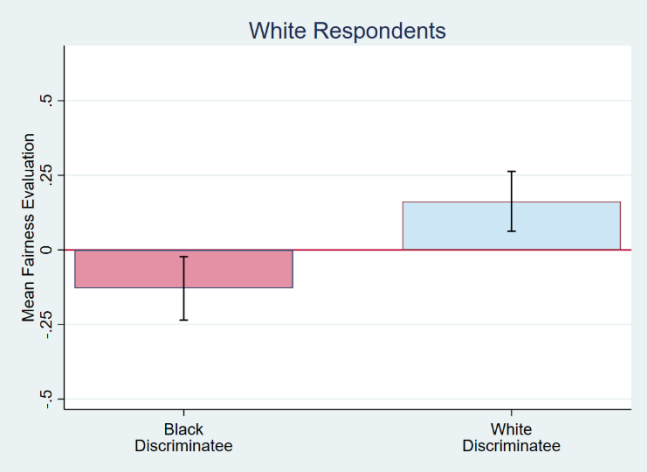
Within Black Discriminatees: *p*=.000

Within White Discriminatees: *p*=.000

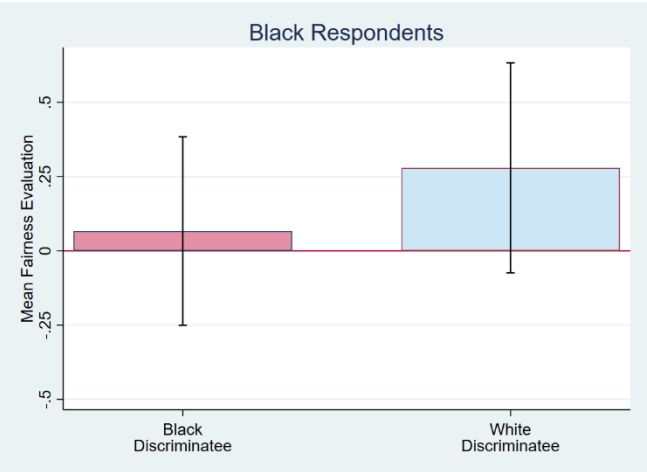
Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.469 standard deviations less fair. Within White Discriminatees, less-justifiable scenarios are 0.495 standard deviations less fair. A test for equality of the Less versus More *Justifiability* Gap between the Black and White treatment yields *p* = .679.

Figure A11.3: Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 4)

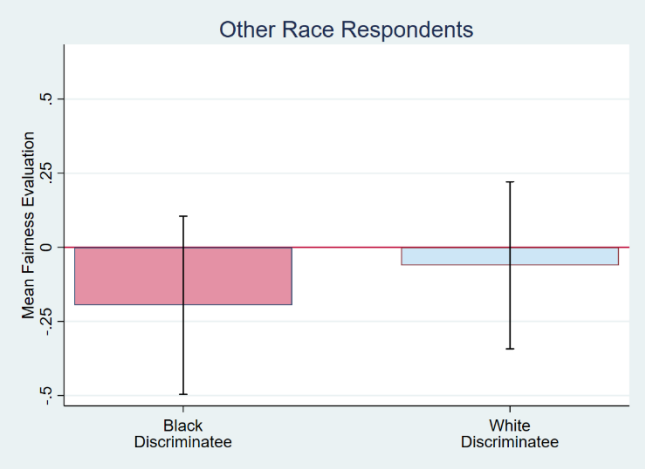
(a)



(b)



(c)



p-values:

Black versus White Treatment:

Overall: *p*=.000

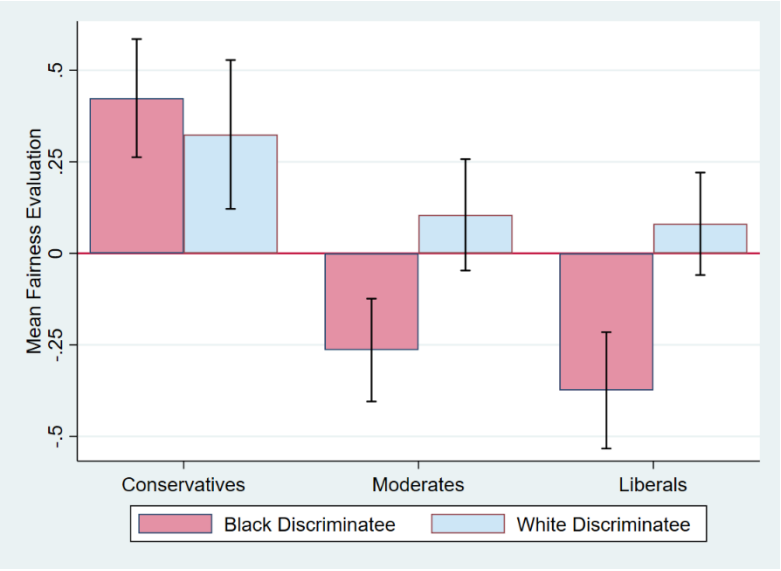
Within White respondents: *p*=.000

Within Black respondents: *p*=.365

Within Other respondents: *p*=.506

Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) across all three racial groups yields *p* = .739.

Figure A11.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Figure 5)



p-values:

Black versus White Treatment:

Overall: *p*=.000

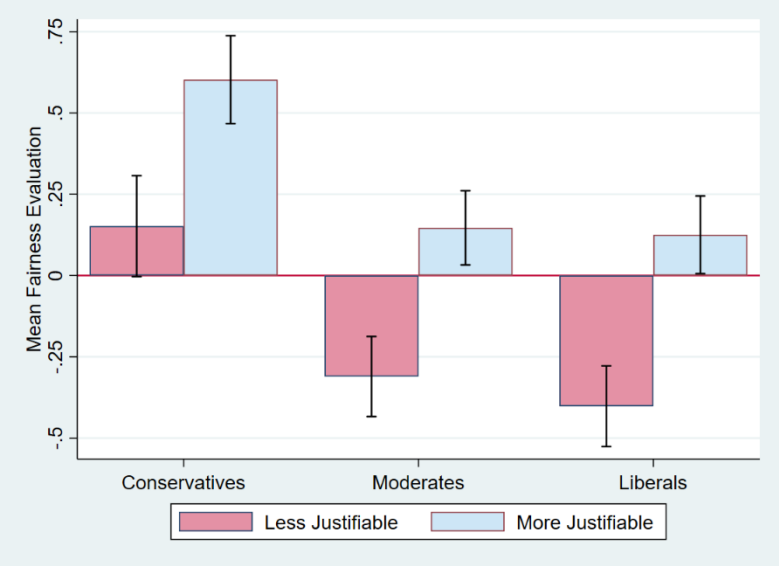
Within Conservatives: *p*=.448

Within Moderates: *p*=.000

Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields $p = .567$. A test for equality between conservatives and (moderates + liberals) yields $p = .001$.

Figure A11.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent's Political Leaning (replicates Figure 6)



p-values:

Less versus More Justifiable Treatment:

Overall: *p*=.000

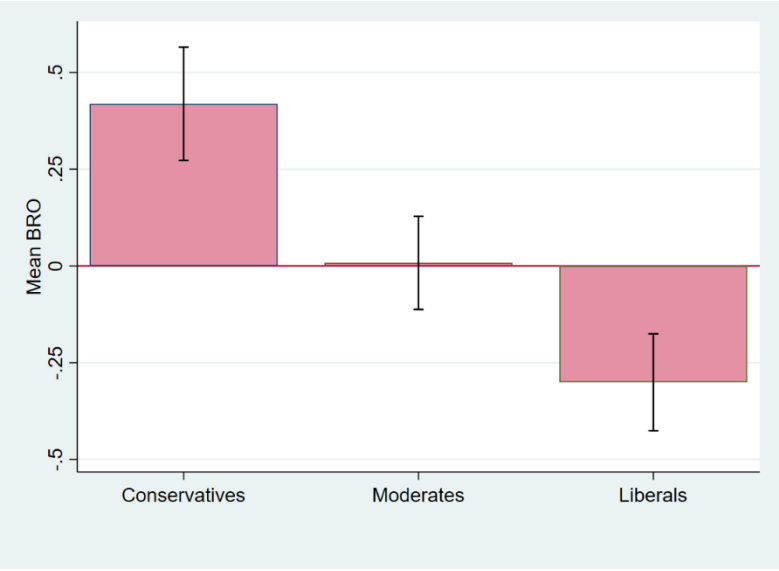
Within Conservatives: *p*=.000

Within Moderates: *p*=.000

Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the Less versus More *Justifiability* Gap across Conservatives, Moderates, and Liberals yields *p* = .590.

Figure A11.6: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 7)



p-values:

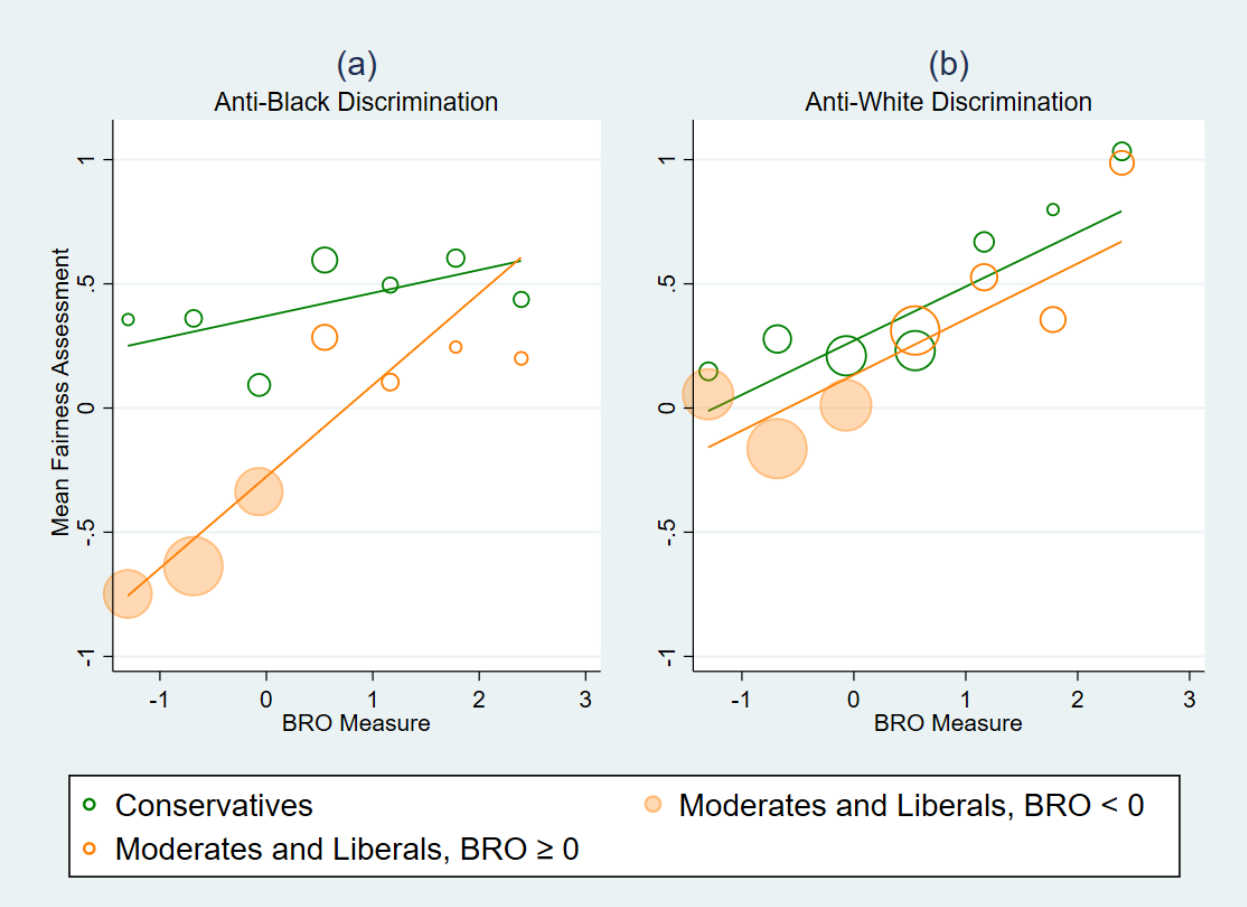
Conservatives vs. Moderates: *p*=.000

Moderates vs. Liberals: *p*=.001

Liberals vs. Conservatives *p*=.000

Notes: BRO is the respondent's assessment of Black peoples' relative economic opportunity, where the raw measure runs on a scale of -3 (much less) to 3 (much more). However, this figure is based on a standardized version of BRO. It is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of BRO across all three political groups yields *p* = .577.

Figure A11.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 8)



Notes: Symbol size is proportional to the number of respondents. Sample is restricted to Stage 1 fairness assessments only. The *p*-values below are clustered by respondent, except for those pertaining to Panel (c).

- Panel (a), Discrimination against Black Applicants
  - For Conservatives: slope = 0.093, p = .218
  - For Moderates and Liberals, slope = 0.369, p = .000
- Panel (b), Discrimination against White Applicants
  - For Conservatives: slope = 0.218, p = .094
  - For Moderates and Liberals, slope = 0.224, p = .000

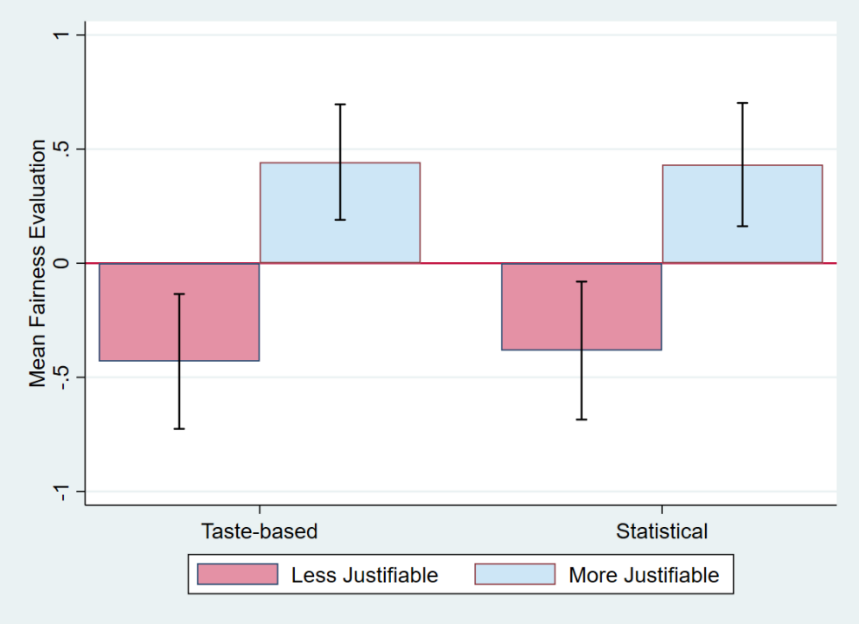
Appendix 12: Replicating the Main Figures for 'Thoughtful' Subjects Only

In this Appendix, we replicate Figures 2-8 with a subsample of "thoughtful" respondents. These respondents took more than the median amount of time (i.e., 8.37 minutes) to read our vignettes and think about their fairness assessments. This sample is composed of approximately 30% of respondents who identify as conservatives, 35% who identify as moderates, and 35% who identify as liberals. A comparison in the demographics between the full sample and subsample of thoughtful respondents is provided below in Table A12.1.

Table A12.1: Composition of MTurk Sample versus "Thoughtful" Subsample

CHARACTERISTIC	Full Sample (1)	"Thoughtful" Sub-sample (2)
Male	0.600	0.553
Female	0.400	0.447
White respondent	0.780	0.750
Black respondent	0.115	0.131
Asian respondent	0.042	0.038
Hispanic respondent	0.037	0.044
American Indigenous respondent	0.009	0.016
Pacific Islander respondent	0.005	0.009
Other race respondent	0.011	0.013
Age 18-24	0.037	0.022
Age 25-34	0.435	0.488
Age 35-44	0.294	0.256
Age 45-54	0.146	0.138
Age 55-64	0.061	0.066
Age 65 and over	0.026	0.031
High School or less	0.098	0.066
2-year or some college	0.196	0.147
4-year college or university	0.519	0.566
Higher degree	0.187	0.223
Observations	642	320

Figure A12.1: Fairness Ratings by Type of Discrimination and *Justifiability* (replicates Figure 2)



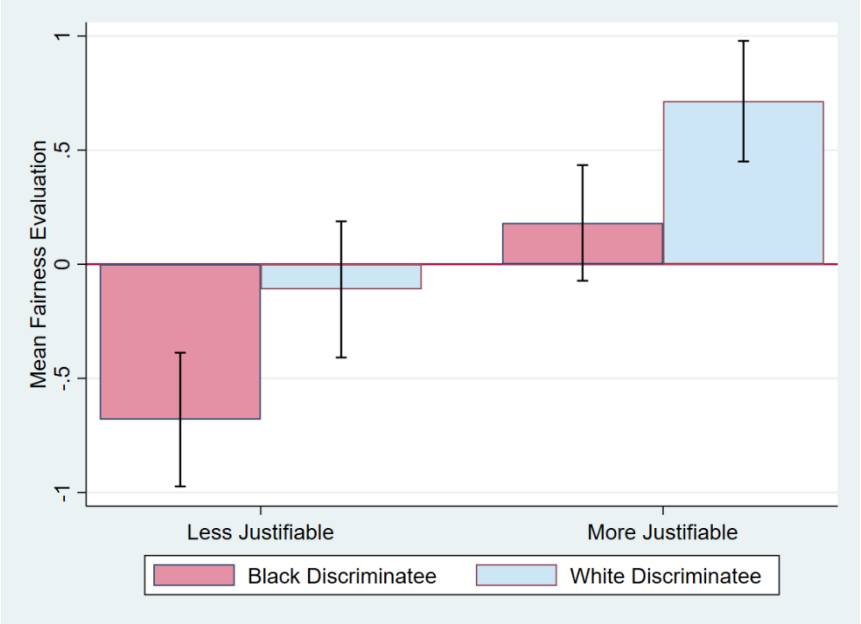
p-values:

Less- versus more justifiable treatments:
Overall: *p*=.000
Within taste-based: *p*=.000
Within statistical: *p*=.000

Taste versus Statistical Discrimination:
Overall: *p*=.918
Within Less-Justifiable: *p*=.824
Within More-Justifiable: *p*=.953

Note: This figure is based on only Stage 1 observations. All p-values are clustered by respondent.

Figure A12.2: Fairness by *Justifiability* and Discriminatee Race (replicates Figure 3)



p-values:

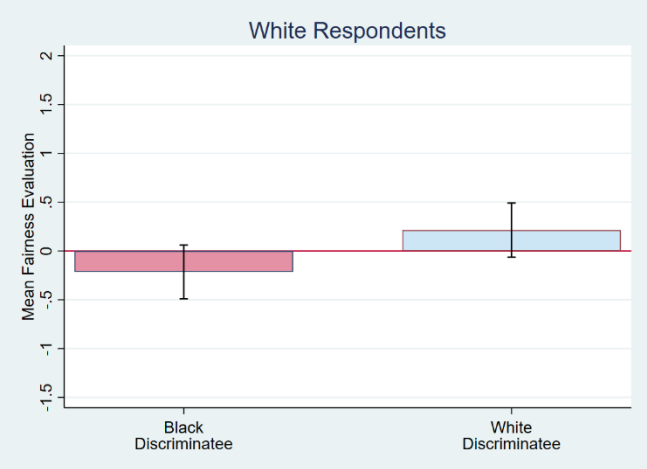
Black versus White Treatment:
Overall: *p*=.002
Within Less-Justifiable: *p*=.007
Within More-Justifiable: *p*=.004

Less versus More Justifiable Treatment:
Overall: *p*=.000
Within Black Discriminatees: *p*=.000
Within White Discriminatees: *p*=.000

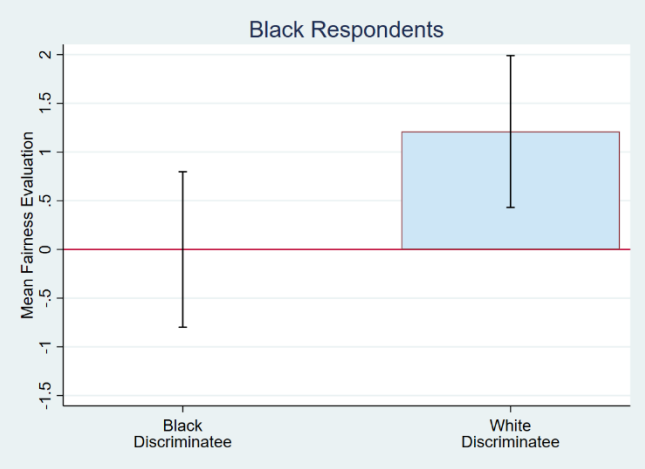
Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.861 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.825 units less fair. A test for equality of the Less versus More *Justifiability* Gap between the Black and White treatment yields *p* = .845.

Figure A12.3: Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 4)

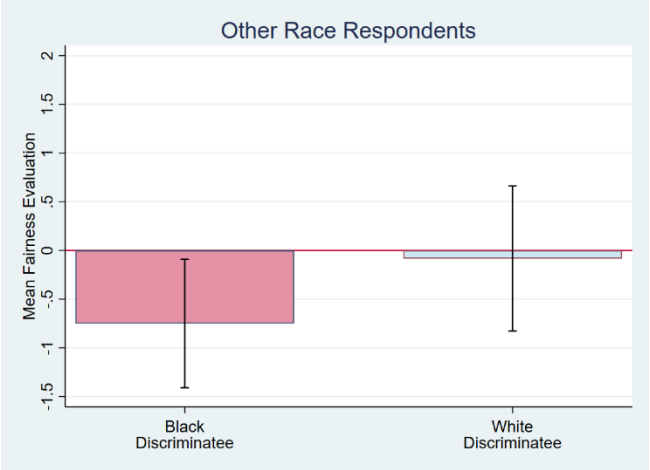
(a)



(b)



(c)

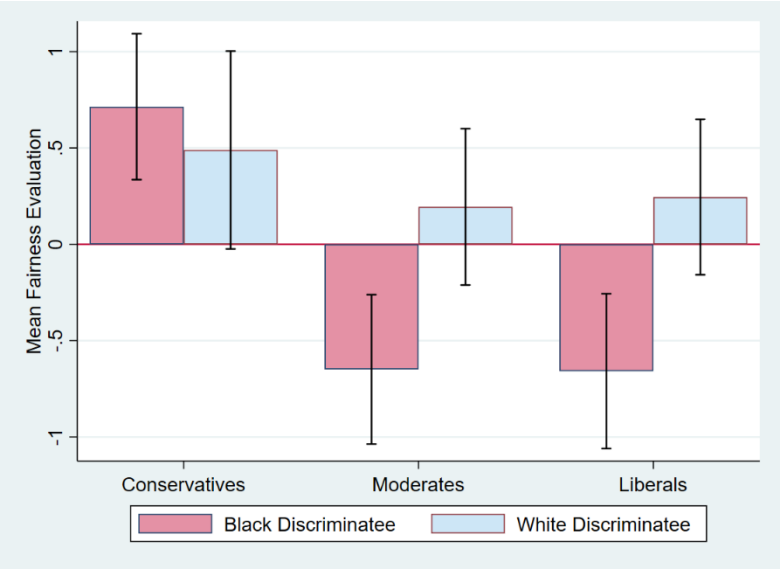


p-values:

Black versus White Treatment:
Overall: Overall: *p*=.000
Within White respondents: *p*=.031
Within Black respondents: *p*=.028
Within Other respondents: *p*=.164

Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e. the Black treatment) across all three racial groups yields *p* = .261.

Figure A12.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Fig. 5)

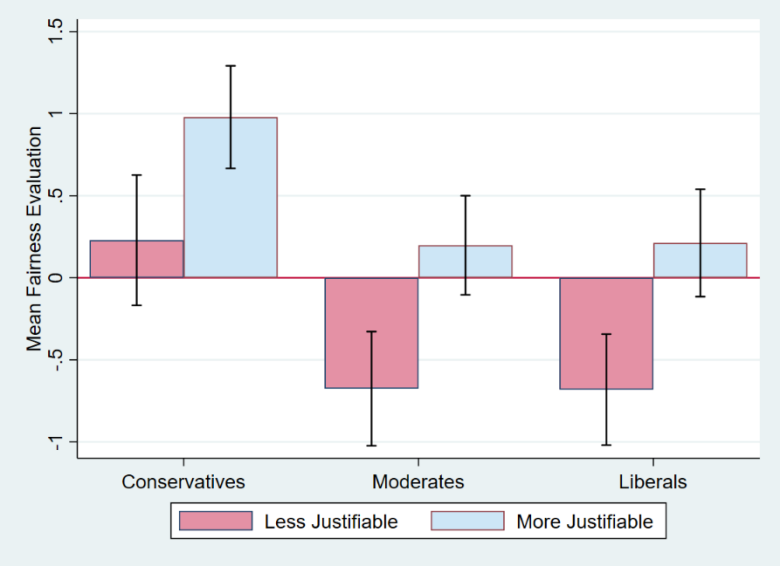


p-values:

Black versus White Treatment:
Overall: *p*=.002
Within Conservatives: *p*=.323
Within Moderates: *p*=.000
Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields *p* = .880. A test for equality between conservatives and (moderates + liberals) yields *p* = .003.

Figure A12.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent's Political Leaning (replicates Fig. 6)

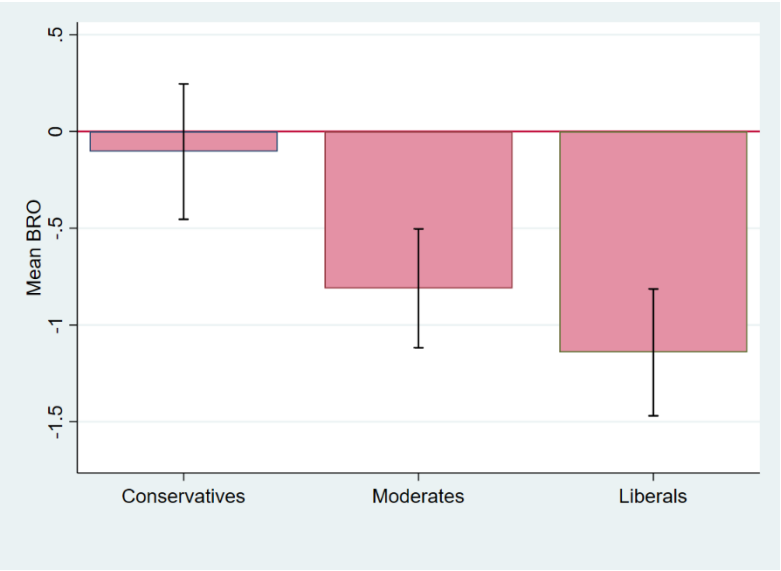


p-values:

Less versus More Justifiable Treatment:
Overall: *p*=.000
Within Conservatives: *p*=.000
Within Moderates: *p*=.000
Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the Less versus More *Justifiability* Gap across Conservatives, Moderates, and Liberals yields *p* = .590.

Figure A12.6: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 7)



p-values:

Conservatives vs. Moderates: *p*=.003

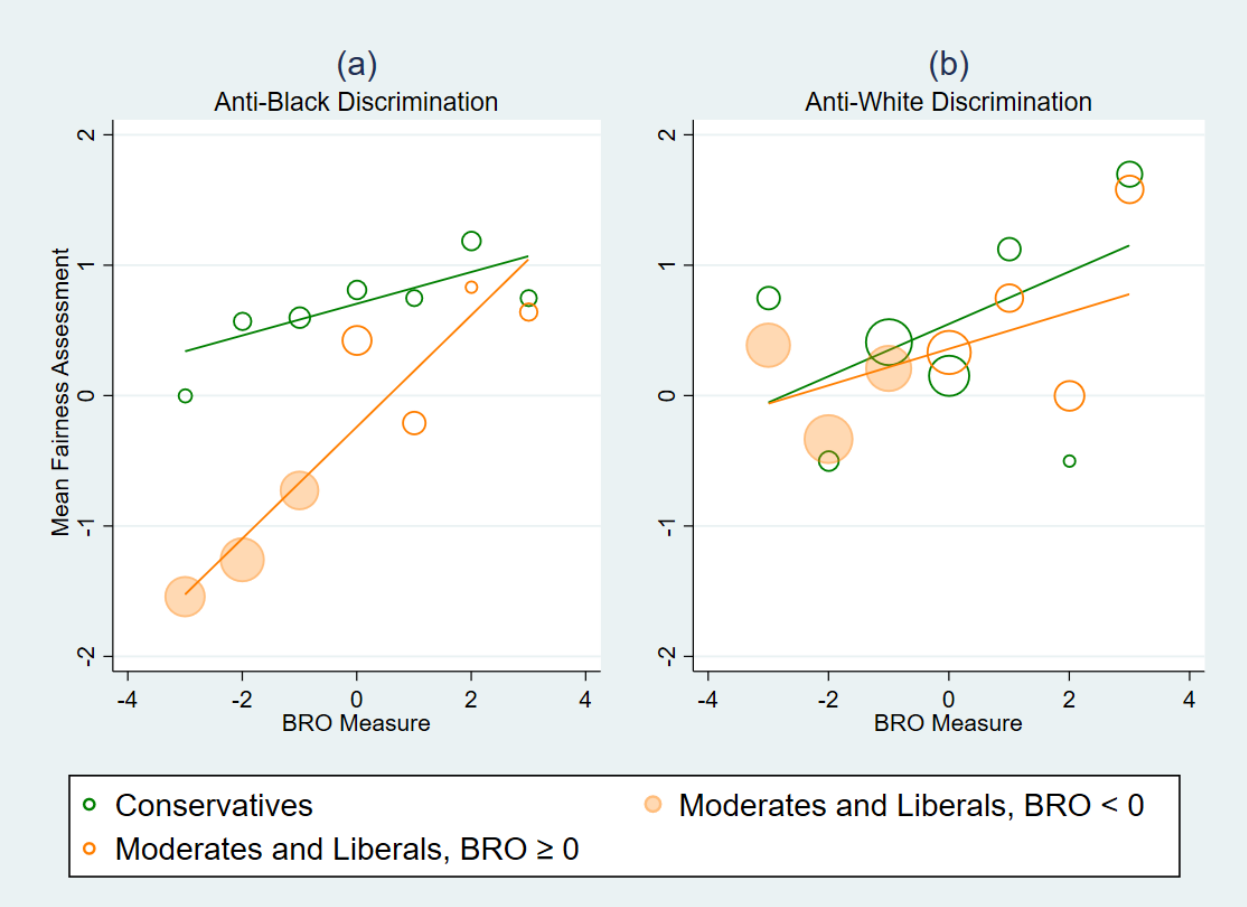
Moderates vs. Liberals: *p*=.145

Liberals vs. Conservatives *p*=.000

Notes:

BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more). This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of BRO across all three political groups yields $p = .672$.

Figure A12.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 8)



Notes: Symbol size is proportional to the number of respondents. Sample is restricted to Stage 1 fairness assessments only. The *p*-values below are clustered by respondent, except for those pertaining to Panel (c).

- Panel (a), Discrimination against Black Applicants
  - For Conservatives: slope = 0.122, $p = .211$
  - For Moderates and Liberals, slope = 0.428, $p = .000$
- Panel (b), Discrimination against White Applicants
  - For Conservatives: slope = 0.201, $p = .281$
  - For Moderates and Liberals, slope = 0.140, $p = .106$

Appendix P: Populated Pre-Analysis Plan

On September 21, 2020, we posted a pre-analysis plan on the AEA RCT Registry.[6] Our experiment was conducted on MTurk in multiple waves between September 22 and October 6, 2020, yielding a final sample of 642 respondents. For each research question in the PAP, this Appendix does two things:

- We present and discuss the results of any exact statistical test or regression analysis that was proposed in the PAP.
- We describe where and how we ultimately addressed that research question in the paper.

Following the PAP (which is downloadable from the AEA Registry), the first three Sections of this Appendix focus on three research questions in turn: establishing the main facts, exploring some simple models of subjective fairness, and robustness/heterogeneity.[7] For easy comparison, all these Sections and sub-Sections are numbered in the same way as the PAP. The final Section of the Appendix summarizes the main similarities and differences between the PAP and the paper.

[6] Our PAP can be downloaded from the AEA RCT Registry under the following entry: Kuhn, Peter and Trevor Osaki. 2020. "When is Discrimination Unfair?." AEA RCT Registry. https://doi.org/10.1257/rct.6409-1.0.

[7] As proposed in the PAP, the fairness measures used within the following analyses are standardized with respect to the full sample.

P1. Establishing the Main Facts

P1.1 Is Taste-Based Discrimination Seen as Less Fair than Statistical Discrimination?

P1.2 How Do People Respond to Sub-types of Taste-Based and Statistical Discrimination?

P1.3 Do People React Differently to Discrimination Against Their Own Race versus Other Races?

PAP items 1.1 -1.3 proposed simple *t*-tests of the above hypotheses, all conducted on the full sample of survey responses, clustering standard errors by respondent. These tests are implemented as univariate regressions in Table P1.1, which shows that:

- Contrary to what we expected from our reading of the economics literature, respondents do not distinguish between scenarios that depict taste-based versus statistical discrimination.
- As hypothesized, respondents object more strongly to taste-based discrimination by employers when it is based on the employer's own tastes (rather than the tastes of his customers).
- As hypothesized, respondents object more strongly to statistical discrimination based on low-quality information, compared to high-quality information.
- Respondents object more strongly to anti-Black than to anti-White discrimination. While the point estimate of this *discriminatee race effect* is similar for White and Non-White respondents, it is not statistically significant in the non-White sample, which is much smaller in size.

Table P1.1: How the type of discrimination, subcases, and respondents' own race affect fairness assessments

	All Respondents (1)	All Respondents (2)	All Respondents (3)	All Respondents (4)	White Respondents (5)	Non-white Respondents (6)
Taste-based	-0.0384					
	(0.0448)					
Taste-based × Employer		-0.474***				
		(0.0354)				
Statistical × Low-quality			-0.490***			
			(0.0388)			
Black discriminatee				-0.181***	-0.190***	-0.145
				(0.0426)	(0.0474)	(0.0962)
Constant	0.0191	0.217***	0.264***	0.0919**	0.0888**	0.103
	(0.0376)	(0.0408)	(0.0405)	(0.0374)	(0.0408)	(0.0884)
Observations	2,568	1,276	1,292	2,568	2,004	564
R-squared	0.000	0.056	0.060	0.008	0.009	0.005

Notes: This table contains the results of parts 1.1-1.3 from the pre-analysis plan. Three stars indicate a one percent significance level. Standard errors are clustered by the respondent.

In the paper, we use similar *t*-tests to compare the fairness of Statistical and Taste-Based Discrimination, as well as the sub-types of each (which we collectively call more- versus less-justifiable discriminatory acts) in Figure 2. The only difference from the PAP is that we restrict the sample to Stage 1 survey responses. This was to avoid possible contamination by the question order effects for the *race* treatment we discovered. The results are essentially identical to the PAP. We explored how the discriminatee race effect varies with the respondent's own race in Figure 4 (which implements a similar *t*-test) and discuss the implications of our findings for the racial in-group bias model in Section 4.2. The in-group bias model is rejected in all cases.

Motivated by the *race* treatment order effects described above, research questions 1.4-1.6 and 2.1–2.4 *all* restrict their analysis to Stage 1 responses when they are addressed in the paper. (Here in the populated PAP we use all responses, as originally specified.) [8]

P1.4 Determinants of Black People's Perceived Relative Opportunities (BRO)

PAP item 1.4 proposed to address the question "How Do Perceptions of Black and White Peoples' Relative Opportunities Vary with Race, Gender, Age, and Political Preferences?" by running the following regression:

$$BRO_i = \alpha + \theta^1 RR_i + \theta^2 RG_i + \theta^3 RA_i + \theta^4 RP_i + \varepsilon_i \qquad (1)$$

where BRO_i is respondent *i*'s assessment of Black peoples' relative opportunities.[9] *RR, RG, RA,* and *RP* represent (sets of) dummy variables for respondent race, gender, age, and political preferences, respectively. The PAP stated that we do not have strong priors for these effects, though we noted that factors like in-group bias could generate motivated beliefs about relative opportunities. The results of this regression are reported in Table P1.4.

According to the Table, respondents' race, gender, and age do not have significant effects on their perceptions of BRO. Democrats, Independents, Liberals, and Moderates all believe that Black people have fewer economic opportunities than Republicans and Conservatives. Finally, as discussed in the paper, the perceived fairness of discriminatory acts *increases* with the respondent's education level.

In the paper, Figure 6 shows the relationship between the respondent's political leaning and BRO (essentially the θ^4 coefficients in equation 1, without the other controls). The results are very similar. Here, as in most of the paper, we use only political orientation (not party preference) to summarize respondents' political stance, in part because independent voters appear to be a more heterogeneous group than self-identified moderates.

[8] Except in the small handful of cases where noted, this sample restriction has no effect on the results.

[9] Due to a cut-and-paste error, the PAP erroneously stated that equation 1 would be estimated using about 2400 fairness assessments (about 600 from each subject). BRO was elicited only once per subject in the survey, however, so the actual regression only contains one observation per respondent.

Table P1.4: How Do Perceptions of Relative Opportunities Vary with Characteristics?

	(1)
Black respondent	-0.00337
	(0.107)
Other race respondent	-0.107
	(0.136)
Male	0.0693
	(0.0784)
Age 35-44	-0.0494
	(0.0915)
Age 45-54	-0.161
	(0.105)
Age 55 and over	-0.122
	(0.144)
Democrat	-0.446***
	(0.102)
Independent or other party	-0.321**
	(0.130)
Liberal	-0.449***
	(0.119)
Moderate	-0.223**
	(0.110)
Four-year college	0.193**
	(0.0849)
Graduate School	0.282**
	(0.117)
Constant	0.363***
	(0.129)
Observations	642
R-squared	0.135

Notes: This table contains the results of estimating equation (1). The outcome variable, BRO, ranges from -3 and 3. Two stars indicate a five percent significance level, and three stars indicate a one percent level.

P1.5 Determinants of the *Discriminatee Race Effect*

PAP item 1.5 addresses the question "How Does Racial Bias in Fairness Assessments vary with Race, Gender, Age, and Political Preferences?" Pooling all respondent races, all treatments, and both stages of the survey we proposed to run the following regression on a sample of about 2400 fairness assessments:

$$FAIR_{ij} = \alpha + \beta^1 T_{ij} + \beta^2 (S_{ij} \times L_{ij}) + \beta^3 (T_{ij} \times E_{ij}) + \delta B_{ij} + \gamma^1 RR_i + \gamma^2 RG_i + \gamma^3 RA_i + \gamma^4 RP_i + \varphi^1 (RR_i \times B_{ij}) + \varphi^2 (RG_i \times B_{ij}) + \varphi^3 (RA_i \times B_{ij}) + \varphi^4 (RP_i \times B_{ij}) + \varepsilon_{ij} \quad (2)$$

where $FAIR_{ij}$ is respondent *i*'s assessment of the fairness of scenario *j*. In equation (2), *S* and *T* are dummies for statistical and taste-based discrimination, and *L* (low quality information) and *E* (employer tastes) are dummies for the sub-types of discrimination that we hypothesize will be viewed more harshly by respondents. Thus, we expect $\beta^2 < 0$ and $\beta^3 < 0$. Together, the β coefficients summarize the effects of the types of discriminatory *actions* described in our vignettes. B_{ij} equals one if the (fictional) discriminatee is Black. Of central interest, the φ coefficients will reveal how the effect of (being randomly exposed to) a Black discriminatee (B_{ij}) varies with the race, gender, age, and political leanings of the survey respondent.

Results from this regression are displayed in Table P1.5. Panel A shows our experimental treatment effects for a respondent with baseline characteristics (in this case White, female, age 18-34, Republican, conservative, 2 years of college or less). Replicating earlier results, it shows that respondents do not distinguish between Taste-Based and Statistical discrimination, but they do care about the sub-types of each. Also, these baseline respondents (who are politically conservative) do not consider the race of the discriminatee when making their fairness assessments. Panel B reproduces other results we have already established: respondent race, gender and age do not affect fairness assessments, but education and political preferences do. Finally, with the exception of an apparently anomalous effect for respondents over age 55, the only respondent characteristic that significantly interacts with the Black experimental treatment is political leaning: As is documented and explored more fully in the paper, liberal and moderate respondents (unlike conservative respondents) rate discrimination against Black job applicants as significantly less fair than (the same act of) discrimination against White applicants.

In the paper, Figure 4 displays the *discriminatee race effect* by *respondent* race (essentially, equation 2's φ^1 coefficient, but without the other controls). As in Table P1.3, we find no significant differences between the racial groups. Figure 5 displays the discriminatee race effect by political orientation (essentially φ^4). As in Table P1.3, we find large differences: conservatives do not consider respondent race but moderates and liberals do.

Table P1.5: How Does Racial Bias in Fairness Assessments vary with Respondent Characteristics?

	coefficient	standard error
A. Treatment Effects:		
Taste-based	-0.0465	(0.0489)
Statistical × Low-quality info	-0.490***	(0.0390)
Taste-based × Customer	-0.474***	(0.0355)
Black discriminatee	0.0336	(0.130)
B. Respondent Characteristics:		
Black respondent	0.0627	(0.125)
Other race respondent	-0.138	(0.122)
Male	0.0272	(0.0764)
Age 35-44	-0.0392	(0.0859)
Age 45-54	-0.0794	(0.108)
Age55 and over	0.0461	(0.135)
Democrat	-0.233***	(0.0858)
Independent or other party	-0.320***	(0.121)
Liberal	-0.190*	(0.104)
Moderate	-0.131	(0.103)
Four-year college or university	0.240***	(0.0841)
Graduate school	0.433***	(0.107)
C. Race Treatment Interactions with Respondent Characteristics:		
Black Discriminatee × Black respondent	0.0301	(0.102)
Black Discriminatee × Other race respondent	0.0844	(0.148)
Black Discriminatee × Male respondent	0.104	(0.0838)
Black Discriminatee × Age35-44	-0.0686	(0.102)
Black Discriminatee × Age45-54	0.0471	(0.111)
Black Discriminatee x Age 55 and over	-0.300**	(0.140)
Black Discriminatee × Democrat	-0.110	(0.0938)
Black Discriminatee × Independent or other party	0.0310	(0.139)
Black Discriminatee × Liberal	-0.270**	(0.114)
Black Discriminatee × Moderate	-0.266**	(0.112)
Black Discriminatee × Four-year college	0.0447	(0.0938)
Black Discriminatee × Graduate School	-0.120	(0.118)
Constant	0.427**	(0.133)
Observations	2,568	
R-squared	0.169	

Note: This table contains the results of estimating equation (2) from the pre-analysis plan. One star indicates a ten percent significance level, two stars indicate a five percent level, and three stars indicate a one percent level. Standard errors are clustered by the respondent.

P1.6 The Relative Importance of "Actions" versus "Identity"

PAP item 1.6 addresses the question "What Matters More for the Perceived Fairness of Discrimination: Actions or Identity?" Here we again pool all respondent races, all treatments, and both stages of the survey to obtain about 2400 evaluations of discriminatory acts from about 600 respondents. In this sample, we run the following regression:

$$FAIR_{ij} = \alpha + \beta^1 T_{ij} + \beta^2 (S_{ij} \times L_{ij}) + \beta^3 (T_{ij} \times E_{ij}) \qquad (3)$$
$$+ \delta^1 RW_i + \delta^2 RB_i + \delta^3 (RW_i \times B_{ij}) + \delta^4 (RO_i \times B_{ij}) + \delta^5 (RB_i \times B_{ij}) + \varepsilon_{ij}$$

As in equation (2), the β coefficients capture the effects of the types of discriminatory *actions* in our survey in the greatest detail possible. The δ coefficients use a relatively expansive set of respondent race categories (White (RW), Black (RB) and Other (RO)), interacted with the Black experimental treatment (B) to capture the effects of racial *identity* on perceived fairness of discrimination.[10]

As laid out in the PAP, Table P1.6 estimates equation (3) three different ways: in its entirety (column 1), then using only the "actions" or "identity" covariates alone (columns 2 and 3). Comparing the regression R^2s, it is clear that actions explain much more of the variation fairness assessments (5.8%) than the identities of the respondent and the (fictitious) discriminatee (1.3%).

While we still think it is of some interest, we chose not to focus on Table P1.6's *actions vs. identity* decomposition ***in the paper***. That said, we note that Table P1.6's results (that actions matter more) are consistent with three of the paper's main findings: (i) that respondents of all political orientations care strongly, and in the same, race-blind way, about the justifiability of actions; (ii) that the *respondent*'s race does not markedly affect fairness assessments; and (iii) that only moderate/liberal respondents care about the race of the (fictional) discriminatee.

[10] As already noted, in most of our analysis we use only two racial categories –White and Non-White—since we do not expect to have enough Black respondents to treat them separately. Here, however, our goal is to absorb as much variation in both actions and racial identity as possible, to see which contributes the most to perceptions of fairness.

Table P1.6: What Matters More – Actions or Identity?

	Actions & Identity (1)	Actions (2)	Identity (3)
Taste-based	-0.0533	-0.0467	
	(0.0488)	(0.0493)	
Statistical × Low-quality	-0.490***	-0.490***	
	(0.0389)	(0.0388)	
Taste × Employer	-0.474***	-0.474***	
	(0.0354)	(0.0354)	
White respondent	0.179		0.178
	(0.122)		(0.122)
Black respondent	0.364**		0.363**
	(0.172)		(0.172)
Black discriminatee × White Respondent	-0.191***		-0.190***
	(0.0476)		(0.0475)
Black discriminatee × Other race respondent	-0.0635		-0.0627
	(0.124)		(0.124)
Black discriminatee × Black respondent	-0.214		-0.217
	(0.146)		(0.146)
Constant	0.178	0.264***	-0.0889
	(0.119)	(0.0405)	(0.115)
Observations	2,568	2,568	2,568
R-squared	0.072	0.058	0.013

Notes: This table contains the results of estimating equation (3) from the pre-analysis plan. Column 1 includes all the covariates of this equation. Column 2 only includes the covariates pertaining to the types of discriminatory scenarios. Finally, Column 3 only includes the covariates pertaining to respondents' racial groups. Two stars indicate a five percent significance level, and three stars indicate a one percent level. Standard errors are clustered by the respondent.

P2. Exploring Some Simple Models of Subjective Fairness

P2.1 The Utilitarian Social Preferences Model

P2.2 The Rules-Based Fairness Model

P2.3 The In-Group Bias Model

In these three parts of the PAP we proposed to explore the potential of three possible models of fairness –utilitarianism, rules-based fairness, and in-group bias-- in accounting for our respondents' fairness assessments. This was done by estimating variations of the following generalized regression model:

$$FAIR_{ij} = \alpha + \beta A_{ij} + \delta B_{ij} + \varepsilon_{ij} \quad (4)$$

where A_{ij} is a set of dummy variables capturing the types and sub-types of discriminatory *actions* that took place in the scenario (e.g. employer-based taste discrimination), and B_{ij} indicates a (randomly assigned) Black discriminatee. Results from these regressions are provided in Table P2.1.

Columns 1 and 2 include all respondents, regardless of their race. They show support for both rules-based fairness (because the sub-types of discrimination matter) and utilitarianism (because anti-Black discrimination is seen as less fair than anti-White discrimination). Columns 3 and 4 restrict attention to White respondents, with similar results. However, the fact that White respondents, as a group, see anti-Black discrimination is seen as less fair than anti-White discrimination is inconsistent with the in-group bias model. Finally, Columns 5 and 6 restrict attention to non-White respondents. Interestingly, while statistical power for this group is lower, the respondent-fixed-effect model suggests that these respondents react to all our experimental treatments (including discriminatee race) the same way. Overall, these results are much more consistent with a model in which White and non-White respondents share similar utilitarian preferences than a model of racial in-group bias.

In the paper, the "utilitarian social preferences model" (now *Utilitarianism*) is tested in Section 4.1. While reject the model for conservative respondents, it is consistent with the response behavior of moderates and liberals. The "rules-based fairness model" (now *Race-Blind Rules*, or *RBRs*) is tested in Section 4.3. In this model, respondents care about the actions that were taken (Tastes vs. Statistical, more- versus less justifiable); further, their valuations of these actions should be invariant to the race of the discriminatee. (For example, if a less-justifiable act is X units less fair than a more-justifiable act against a White discriminatee, the same fairness penalty should apply to a Black discriminatee). We find strong support for this model for respondents of all political leanings. Finally, the "in-group bias model" (now labeled more precisely as *racial in-group bias*) is tested in Section 4.2. Our statistical power is too low to draw conclusions for non-White respondents, but (as in the PAP) we decisively reject it for White respondents.

Table P2.1: Assessing Three Models of Fairness

	All Respondents (1)	All Respondents (2)	White Respondents (3)	White Respondents (4)	Non-White Respondents (5)	Non-White Respondents (6)
Taste-based	-0.0498	-0.0108	-0.0833	0.00662	0.0669	-0.121
	(0.0492)	(0.0513)	(0.0559)	(0.0592)	(0.103)	(0.197)
Statistical × Low-quality	-0.490***	-0.490***	-0.531***	-0.531***	-0.344***	-0.660***
	(0.0388)	(0.0449)	(0.0440)	(0.0508)	(0.0822)	(0.182)
Taste × Employer	-0.474***	-0.474***	-0.462***	-0.462***	-0.514***	-0.986***
	(0.0354)	(0.0408)	(0.0403)	(0.0465)	(0.0742)	(0.165)
Black discriminatee	-0.181***	-0.163***	-0.191***	-0.151***	-0.145	-0.374**
	(0.0427)	(0.0378)	(0.0476)	(0.0430)	(0.0964)	(0.154)
Constant	0.358***	0.0862**	0.379***	0.0787*	0.283***	2.159***
	(0.0459)	(0.0369)	(0.0502)	(0.0425)	(0.108)	(0.141)
Observations	2,568	2,568	2,004	2,004	564	564
R-squared	0.067	0.695	0.074	0.687	0.047	0.725
Respondent FE	NO	YES	NO	YES	NO	YES

Notes: This table contains the results of estimating equation (4) from the pre-analysis plan. Columns 1-2 include all respondents, regardless of their race. Columns 3-4 only include White respondents. Finally, Columns 5-6 only include Non-white respondents. Two stars indicate a five percent significance level, and three stars indicate a one percent level. Standard errors are clustered by the respondent.

P2.4 A Hybrid Model: Conditional Utilitarianism

In this part of the PAP we explore the potential for a conditional utilitarianism model (where different beliefs about relative opportunities explain different discriminatee race effects). Separately for White and Black respondents, we divide respondents into two groups: those who believe Black people have fewer economic opportunities (BFO), and those who believe that Black people have the same or more opportunities (BMO).[11] We then expand equation (4) to include interactions between the Black treatment (*B*, where the discriminatee is Black) and BMO, as follows:

$$FAIR_{ij} = \alpha + \beta A_{ij} + \delta^1 BMO_i + \delta^2 (BFO_i \times B_{ij}) + \delta^3 (BMO_i \times B_{ij}) + \varepsilon_{ij} \quad (5)$$

In equation (5), δ^1 measures the extent to which discrimination against White people (the omitted discriminatee category) is more acceptable among respondents who believe that Black people have more economic opportunities than among respondents with the opposite belief. If our respondents are conditional utilitarians– i.e. they are less tolerant of discrimination against people whom they *believe* have fewer opportunities (who are *White* in this case)—we should see $\delta^1 < 0$. Under the conditional utilitarian model we should also see that people who believe that Black people have fewer opportunities (BFO=1) react more negatively to discrimination against Black people than against White people ($\delta^2 < 0$). Similarly, people who believe that Black people have more opportunities should react less negatively to discrimination against Black people than against White people ($\delta^3 > 0$).

Table P2.4 contains our estimates of equation (5). Consistent with conditional utilitarianism, we find that $\delta^2 < 0$: People who believe that Black people have fewer opportunities (BFO=1) react more negatively to discrimination against Black people than against White people. Inconsistent with that $\delta^3 = 0$ and $\delta^1 > 0$. The latter result is especially large in magnitude and statistical significance; it shows that discrimination against White people becomes *more* acceptable as White people's perceived relative opportunities fall (i.e. as BRO rises). This is the opposite of what a conditional utilitarian model predicts.

[11] Thus, BMO = 1 if the respondent chooses responses 4-7 on the raw seven-point BRO (Black relative opportunity scale). BFO=1 for responses 1-3. We combine the equal opportunities category with strictly greater perceived opportunities because we expect the latter group to be considerably smaller in size. We have explored other cut-offs as well, with similar results.

Table P2.4: Testing the Conditional Utilitarianism Model

	All respondents (1)	White respondents (2)	Non-White respondents (3)
Taste-based	-0.0436	-0.0791	0.0777
	(0.0479)	(0.0550)	(0.0942)
Statistical × Low-quality	-0.490***	-0.531***	-0.344***
	(0.0389)	(0.0440)	(0.0824)
Taste × Employer	-0.474***	-0.462***	-0.514***
	(0.0354)	(0.0403)	(0.0744)
BMO (δ^1)	0.445***	0.336***	0.801***
	(0.0756)	(0.0847)	(0.164)
BFO × Black discriminatee (δ^2)	-0.312***	-0.347***	-0.199*
	(0.0502)	(0.0560)	(0.110)
BMO × Black discriminatee (δ^3)	0.0219	0.0511	-0.0463
	(0.0630)	(0.0709)	(0.146)
Constant	0.193***	0.256***	-0.0182
	(0.0543)	(0.0599)	(0.122)
Observations	2,568	2,004	564
R-squared	0.162	0.155	0.208

Note: This table contains the results of estimating equation (5) from the pre-analysis plan. Columns 1 includes all respondents, regardless of their race. Columns 2 only includes White respondents. Finally, Columns 3 only includes Non-white respondents. One star indicates a ten percent significance level, two stars indicate a five percent level, and three stars indicate a one percent level. Standard errors are clustered by the respondent.

P2.5 Interactions between Distributional Considerations and Concerns for Procedural Fairness

PAP item 2.5 explores whether we might be able to *leverage the within-subject component* of our experimental design to study how subjects' preferences for race-blind rules interact with their utilitarian preferences when those preferences conflict, i.e. when a respondent encounters a change in the Race treatment. The idea is to introduce respondent fixed effects to equation (4) to generate purely *within-subject* estimates of seeing a Black discriminatee (δ). If these effects are smaller in magnitude than the estimates in equation (4) –and especially if they are smaller than purely *between-subject* estimates of δ from stage 1 of the survey only—this would suggest that subjects care about race-blindness by trying to treat discriminatees of the same race the same way.

To that end, Table P2.5 replicates column 1 of Table P1.1 in three new ways. First, column 2 adds respondent fixed effects, giving us a purely *within-subject* estimate of our experimental treatment effects. Column 3 contains estimates from a sample with only Stage 1 observations. Since there is no within-subject variation in the Black treatment during Stage 1, this gives us a purely *between-subject* estimate of that treatment's effects. Finally, Column 4 is estimated using only Stage 2 observations. These estimates are also between-subject, but they may be influenced framing effects related to the treatment the subject encountered in Stage 1.

While the estimates of the Taste, Statistical x Low-quality, and Taste x High-quality treatments are essentially identical across all the columns of Table P2.5, the estimates of the Black treatment tell an intriguing story: The 'pure' *between-subject* estimate of the Black treatment effect (-.505) is considerably larger than all the other estimates. The pure *within-subject* estimate is lower than the overall estimate, and the between-subject Stage 2 estimate is indistinguishable from zero. While this evidence is only suggestive, it suggests that respondents who have experienced a switch in their Race treatment may moderate their Stage-2 fairness assessments in the direction of race-blindness. Inspired by these results from the PAP, we explore treatment order effects in more detail in the main paper and argue that they can provide some insights into how liberals and moderates –the only respondents who care about both utilitarianism and race-blindness—reconcile those objectives when they conflict.

Less formally, the PAP proposes going beyond the comparisons summarized in Table P2.5 by "leverag[ing] the within-subject component of our experimental design to study how subjects' concerns for procedural fairness ('a consistent set of rules for everyone') might interact with their concerns for outcomes, whether driven by bias or utilitarianism." We provided the following illustration of the interactions we had in mind:

> "For example, in-group-biased White respondents who are very tolerant of discrimination against Black people in stage 1 of the experiment might feel the need to be similarly tolerant of discrimination against White people in stage 2, if they care about rules-based ethics as well as outcomes. More generally, a certain form of order effects—specifically, where the discriminatee race a subject is exposed to in the first stage affects their second-stage fairness ratings—would be evidence that subjects are trying to treat the same situation the same way, regardless of the participants' identities."

In the paper, treatment order effects resembling the ones described above are documented in Section 2.4. We then push further on this idea in Section 5, where we first document that these order effects are only present among moderate and liberals, and that they cannot easily be explained by experimenter demand effects. Finally, we interpret these order effects as driven by moderates' and liberals' desires to reconcile the two fairness criteria they care about –utilitarianism and race-blind rules-- when those criteria conflict. We estimate that moderates and liberals place roughly equal weight on these two criteria when they are forced to choose between them.

Table P2.5: Leveraging Within-Subject Treatment Variation to Learn About Preferences for Race-Blindness

	Full Sample	Within-subject	Stage 1 (Between-subject)	Stage 2
	(1)	(2)	(3)	(4)
Taste-based	-0.0956	-0.0207	-0.0555	-0.126
	(0.0944)	(0.0984)	(0.140)	(0.141)
Statistical × Low-quality	-0.941***	-0.941***	-0.970***	-0.909***
	(0.0746)	(0.0861)	(0.0983)	(0.0965)
Taste-based × Employer	-0.909***	-0.909***	-0.875***	-0.940***
	(0.0679)	(0.0784)	(0.0883)	(0.0865)
Black discriminatee	-0.348***	-0.313***	-0.505***	-0.192
	(0.0820)	(0.0726)	(0.129)	(0.133)
Constant	4.401***	3.880***	4.492***	4.306***
	(0.0881)	(0.0707)	(0.114)	(0.125)
Observations	2,568	2,568	1,284	1,284
R-squared	0.067	0.695	0.076	0.062
Respondent FE	NO	YES	NO	NO

Note: Column 1 of Table W2.5 reproduces column 1 of Table W2.1. The remaining columns explore changes to the specification, including adding respondent fixed effects and using data from only one Stage of the experiment. One star indicates a ten percent significance level, two stars indicate a five percent level, and three stars indicate a one percent level. Standard errors are clustered by the respondent.

P3. Robustness and Heterogeneity

P3.1 Heterogeneity

In the PAP, we said we would consider two main types of heterogeneity analysis. The first was to use within-subject estimates of our treatment effects to classify individual respondents into 'types'. We recognized that we should expect very limited statistical power for this exercise, and provided only one example of this idea: using within-subject variation to first identify a set of in-group biased White respondents, then comparing the demographics of this group to the broader population in order to learn "*which* White people exhibit in-group bias?" Due to a combination of limited statistical power and the fact that we found very little evidence of in-group bias, we did not pursue this idea in the paper. As noted below, however, our analysis of heterogeneity on observables found *weak* evidence consistent with racial in-group bias among White conservatives.

The second proposed approach to heterogeneity analysis was to divide the respondents into large sub-samples based on observables, replicating our main analysis by group. The sample divisions we identified as potentially interesting were:

- White, Non-White and Black people
- a small number of respondent Age groups
- men versus women
- college versus non-college-educated respondents
- Republican versus Democrat-leaning respondents

As noted in the paper, we have very limited statistical power for non-White respondents and we do not find strong effects of age or gender on subjects' fairness assessments, so we did not conduct extensive heterogeneity analyses (of treatment effects) on these dimensions. Appendix 2.4 conducts extensive heterogeneity analysis by education and finds that –despite the fact that fairness assessments rise with education overall—all the main treatment effects in our experiment are highly stable across education groups. We interpret this as a difference in fairness 'set points' between education groups. Heterogeneity by political preferences is a central theme throughout the paper, though (as noted) we chose to focus on our indicator of conservative-liberal leaning rather than party preference because Independents could not be easily characterized. For some analyses, we also combined moderates and liberals because their response patterns were so similar. Choices like these are anticipated in the PAP, which stated:

> "We have two indicators of political preference: party preference and a liberal-conservative score. If these are highly correlated (as we expect) we may only use one of them. Another approach might be to reduce the number of categories by allocating conservative persons with Independent party affiliations to the Republican group and liberal Independents to the Democratic group."

P3.2 Robustness

In the PAP we proposed to use standardized (mean 0, standard deviation 1) fairness assessments as our main outcome variables. We abandoned this approach when we realized that our fairness questions contain important cardinal information that would be discarded by such an approach. For example, it matters whether a respondent said discrimination was "very unfair", regardless of how common such assessments were. Thus, all our analyses code "neither fair nor unfair" as a zero, and code (for example) "somewhat fair", "fair" and "very fair" as 1, 2 and 3 respectively. In consequence, our proposed robustness checks for using alternative standardizations (for example allowing individual survey respondents to have a different response variance) is no longer relevant.

In the PAP we proposed some regression analyses that dichotomized the BRO measure and recommended trying alternative cut points for the dichotomization. We now use a continuous version of BRO in Figure 7 so this is no longer relevant either. We also proposed working with more detailed racial identity categories, but (as expected) our samples were much too small for this.

Finally, we proposed to explore if the results change when we restrict attention to more 'thoughtful' subjects who took more time to think about their fairness assessments. We did this in the populated PAP, where Table P3.1 replicates columns 1 and 2 of Table P2.1 ("Assessing Three Models of Fairness") for a subset of respondents who took more than the median amount of time to complete the survey. We also did this in the paper (Appendix 12). In both cases the results were very similar to the entire sample.

Table P3.1: A Look at "Thoughtful" Respondents

	Full Sample (1)	Full Sample (2)	"Thoughtful" Sample (3)	"Thoughtful" Sample (4)
Taste-based	-0.0498	-0.0108	-0.0618	-0.0175
	(0.0492)	(0.0513)	(0.0676)	(0.0730)
Statistical × Low-quality	-0.490***	-0.490***	-0.398***	-0.398***
	(0.0388)	(0.0449)	(0.0512)	(0.0592)
Taste × Employer	-0.474***	-0.474***	-0.440***	-0.440***
	(0.0354)	(0.0408)	(0.0505)	(0.0583)
Black discriminatee	-0.181***	-0.163***	-0.234***	-0.129**
	(0.0427)	(0.0378)	(0.0592)	(0.0568)
Constant	0.358***	0.0862**	0.508***	0.235***
	(0.0459)	(0.0369)	(0.0632)	(0.0757)
Observations	2,568	2,568	1,280	1,280
R-squared	0.067	0.695	0.063	0.671
Respondent FE?	NO	YES	NO	YES

Notes: This table compares estimates for equation (4) between the full sample and a subsample containing respondents that took above the median amount of time to complete the survey on MTurk (i.e., at least 8.5 minutes). These respondents could be relatively more thoughtful than their counterparts. Columns 1-2 contains the estimates for the full sample while 3-4 contains those for "thoughtful" respondents. One star indicates a ten percent significant level, two stars indicate a five percent level, and three stars indicate a one percent level. Standard errors are clustered by the respond

P4. Summary: Comparing the PAP and the paper

P4.1 Key Results in the Paper that were specified in the PAP

- All the descriptive "facts" presented in Section 3.
- All four theoretical models of discrimination described in Section 4, and the main tests thereof. (The models' names have changed slightly.)
- The possibility of question order effects –especially for the *race* treatment--, and the idea of using them to learn about respondents' preferences for race-blindness. (See Appendix P2.5)

P4.2 Main Departures from the PAP in the paper

- Throughout the paper, for simplicity and transparency we decided mostly to report simple *t*-tests of differences in means rather than regression results. In all cases where this is done, the results are extremely similar (in part due to random assignment of treatment).
- While the PAP proposed using standardized (mean 0, standard deviation 1) measures of fairness as our main outcome variables, we realized that this would obscure important cardinal information about levels of fairness. Therefore, we decided to use the raw fairness scores, centered at 0 (corresponding to "neither fair nor unfair").
- Motivated by the *race* treatment order effects, we restricted the sample in Sections 3 and 4 to Stage 1 responses only.
- While we anticipated race treatment order effects, we did not anticipate they would differ by political orientation. We use this distinction in the paper to understand the differences in implicit fairness models between political groups.
- In Figure 8's exploration of the "BRO hypothesis" we decided to use a continuous version of BRO (all seven values) rather than a dichotomized version, to show additional detail.

P4.3 PAP Hypothesis Tests not Included in the Main Paper

- In the PAP, we proposed an "actions versus identity" decomposition. We have performed this decomposition and reported the results in Appendix P1.6, where we also discuss why it did not seem of sufficient interest to include in the main part of the paper.
- Due to a lack of statistical power, we were not able to pursue P3.1's idea of using within-subject variation in responses to treatments to classify subjects into types.