When is Discrimination Unfair?

February 16, 2023

Peter Kuhn[1]

Trevor Osaki[2]

Abstract. Using a vignette-based survey experiment on Amazon's Mechanical Turk, we measure how people's assessments of the fairness of race-based hiring decisions vary with the motivation and circumstances surrounding the discriminatory act and the races of the parties involved. Regardless of their political leaning, our subjects react in very similar ways to the employer's *motivations* for the action, such as the quality of information on which statistical discrimination is based. Compared to conservatives, moderates and liberals are much less accepting of discriminatory actions, and consider the discriminatee's race when making their fairness assessments. We describe four pre-registered models of fairness – (simple) utilitarianism, race-blind rules (RBRs), racial in-group bias, and belief-based utilitarianism (BBU) – and show that the latter two are inconsistent with major aggregate patterns in our data. Instead, we argue that a two-group framework, in which one group (mostly self-described conservatives) values employers' decision rights and the remaining respondents value utilitarian concerns, explains our main findings well. In this model, both groups also value applying a consistent set of fairness rules in a race-blind manner.

---

1. Corresponding author: peter.kuhn@ucsb.edu. UC Santa Barbara, NBER, IZA and CESifo. 2. tosaki@ucsb.edu. UC Santa Barbara. The authors thank Catherine Weinberger, participants in the 2022 Trans-Pacific Labor Seminar and seminar participants at Drexel University, Princeton University, and the University of Georgia for helpful comments. This study and a pre-analysis plan were pre-registered in the AEA RCT Registry, under ID number AEARCTR-0006409.

1. Introduction

A large literature has studied the prevalence, magnitude, and causes of discrimination based on characteristics that include race and gender (Bertrand and Duflo, 2017). Another rapidly growing literature has studied the conditions under which people perceive income and pay inequality as fair or unfair, and has demonstrated that these fairness perceptions can have strong effects on peoples' economic behavior and support for public policies (Alesina and La Ferrara, 2005; Lefgren et al., 2016; Almas et al., 2020; Dube et al. 2021). Motivated by both these literatures, this paper studies whether and when people perceive *discrimination* as unfair—a question that has received much less attention.

To study this question, we use a vignette-based survey experiment on Amazon's Mechanical Turk (MTurk) to measure people's assessments of the fairness of race-based hiring decisions. The vignettes illustrate canonical examples of statistical and taste-based discrimination, with both Black and White recipients of discrimination (*discriminatees*). In addition, the scenarios have varying levels of *justifiability*, i.e., varying motivations for the discriminatory act which we expect will make the actions more or less socially acceptable. The goals of our analysis are, first, to measure the effects of three types of factors on the perceived fairness of a discriminatory act in a broad sample of Americans: the characteristics of the respondent; the motivation for discrimination (e.g., tastes versus statistical); and the identity of the discriminatee (Black versus White). Second, we assess the consistency of four pre-registered models of perceived fairness with the patterns we observe. Finally, we provide a simple, non-preregistered, two-group interpretive framework that provides a convenient summary of all our empirical results.

Our main findings are as follows. First, subjects' self-identified political leanings have large effects on their overall acceptance of discriminatory actions, with conservatives being much more accepting of the discriminatory actions we depict than moderates and liberals. Second, regardless of their political leanings, our respondents care about the detailed motivations behind a discriminatory action (holding the act's consequences constant). Specifically, while the presence of taste-related versus statistical factors does not reliably predict subjects' fairness assessments, other aspects of the discriminator's motivations have robust and sizable effects. For example, discrimination by employers is

seen as substantially less fair when it is based on the employer's own tastes than on the tastes of the employer's customers. Similarly, statistical discrimination is seen as less fair when it is based on low-quality information about relative group productivity, compared to higher-quality information. Notably, the effects of these motivational factors on perceived fairness are very similar across all political groups, and the effects do not depend on the race of the discriminatee.

Third, our moderate and liberal respondents exhibit a strong *discriminatee race effect*: they disapprove more of anti-Black than anti-White discrimination. This effect is absent among conservatives, who rate the discriminatory acts we depict as slightly more fair than unfair, regardless of the discriminatee's race. Fourth, among the four models of perceived fairness we evaluate –(simple) utilitarianism, race-blind rules (RBRs), racial in-group bias, and belief-based utilitarianism (BBU)– the latter two are inconsistent with some major empirical patterns in our data. Fifth, all three political groups in our sample –liberals, moderates, and conservatives—exhibit a strong desire to apply race-blind rules when comparing the fairness of different discriminatory actions. Only liberals and moderates, however, exhibit utilitarian preferences (which assign higher fairness ratings to actions that shift income from high- to low-income groups).

Sixth, to make sense of all these findings, we propose an interpretive framework with two equally sized groups of respondents who collectively care about three fairness criteria: race-blind procedural fairness (RBRs), utilitarianism, and a (non-preregistered) ethic that values employers' decision rights. Both of our groups strongly value race-blind procedural fairness. In addition to this, Group 1, or *Business Rights Advocates,* also care about employer decision rights, but place no value on utilitarian objectives. Group 2, or *Utilitarians,* value utilitarian objectives but exhibit no detectable support for employers' decision rights. Group 1 are predominantly (but not exclusively) self-identified conservatives, while Group 2 is a large subset of moderates and liberals.

Finally, we notice that – unlike Group1 – Group 2 subjects (who are all moderate or liberal) could face a conflict between the two fairness criteria (utilitarianism and race-blind rules) they care about: Objecting more strongly to anti-Black than anti-White discrimination for utilitarian reasons may not feel race-blind. To assess how Group 2 makes this tradeoff, we leverage the fact that a large share of our

subjects experiences a switch in the race of the person being discriminated against during the experiment. Under the assumption that subjects only become aware of their desire to make race-blind fairness assessments after they are exposed to a discriminatee of a second race, we are able to use random assignment of our *race* treatments to estimate that Group 2's fairness assessments place roughly equal weight on these two criteria when they conflict.

Our paper connects to a literature in labor and personnel economics that uses models of fairness to interpret the effects of pay inequality on effort, job performance and satisfaction, wage satisfaction, and quits (Charness and Kuhn 2007; Abeler et al. 2010; Card et al. 2012; Charness et al. 2015; Bracha et al. 2015; Cohn et al. 2015; Breza et al. 2017; Cullen and Perez-Truglia 2018; Dube et al. 2019; Fehr et al. 2021, Schildberg-Hörisch et al. 2022). Some of these authors have argued, for example, that effort- and productivity-related wage differentials are seen as fairer than differentials attributed to other factors, such as luck (Abeler et al., 2010; Breza et al., 2017). We also connect to a literature in experimental and personnel economics on the effects of the intentions behind an economic action on its perceived fairness (Charness and Levine 2000; Offerman 2002; Abeler et al. 2010; Breza et al. 2017). In a variety of contexts, including layoffs and within-firm pay inequality, these authors show that people's reactions to the same action vary dramatically with the reasons why the action was taken. None of these authors, however, consider the effects of the intentions behind a *discriminatory* act on its perceived fairness.[1]

A related literature in sociology has studied peoples' assessments of the fairness of income differentials, in many cases focusing on income gaps between women and men (Jasso and Rossi 1977; Auspurg, Hinz, and Sauer 2017; Jasso, Shelly and Webster 2019; Sauer 2020). Like us, these studies consider a number of implicit criteria people might use to judge the fairness of income differentials; these

[1] In fact, we are aware of only one other study that elicits peoples' assessments of the fairness of discriminatory acts: Feess et al. (2021) use vignettes similar to ours to compare subjects' views of anti-female versus anti-male discrimination. Barr, Lane, and Nosenzo (2018) use an allocator-game lab experiment to elicit second-order beliefs (which discriminatory acts do *others* see as fair?) of British university students. Our focus on first-order beliefs is motivated, in part, by the high level of political polarization in the United States. In such contexts --where social norms are contested—there could be large differences between first- and second-order perceptions of fairness, with the latter being highly sensitive to the identity of the persons whose beliefs the subjects are asked to predict.

criteria include *need* and *impartiality*, which roughly map into our utilitarian and RBR models. To our knowledge, however, this literature has not considered the perceived fairness of discriminatory actions.[2]

Our research also relates to some recent papers that study the effects of peoples' beliefs about the *causes* of inequality on their support for policies that redistribute income and opportunities, both overall (Alesina et al., 2020) and specifically on racial basis (Haaland and Roth 2021; Alesina et al. 2021). The latter two papers find that beliefs about the causes of racial inequality are highly correlated with support for race-based policies like affirmative action; these beliefs also account for much of the partisan divide in policy support. Informational treatments designed to change people's beliefs, however, have limited effects on policy support. Our paper differs from these three papers in two main ways; the first is that we study a different outcome. Specifically, we focus on how our respondents assess the fairness of discriminatory *actions* taken by private individuals (employers in our case), not on respondents' expressions of support for public policies. Second, we consider a broader set of implicit fairness models that people might use to assess either actions or policies. Specifically, we show that peoples' fairness assessments depend not only on an action's consequences (implicit in utilitarian assessments of public policies) but also on the actor's *intentions.* Intentions, and *rules* –i.e. a desire to apply a consistent set of rules when mapping intentions and actions into fairness levels – play important roles in non-consequentialist ethics such as those studied by Andreoni et al. (2019). In our paper we show that expanding the set of fairness models to include these considerations provides a more complete accounting of which types of discriminatory acts (and potentially which types of race-relevant public policies) are perceived as fair or unfair.[3]

---

[2] One recent sociology paper studies how peoples' willingness to engage in (hypothetical) acts of statistical discrimination can be manipulated. Tilcsika (2021) finds that exposing subjects with managerial experience to the theory of statistical discrimination increased the extent to which they relied on gender in a hiring simulation.

[3] Considering non-consequentialist factors may also provide a more complete accounting of which public policies are seen as fair. For example, a restrictive immigration policy might be seen as more fair if it was perceived to be motivated by a sincere desire to protect the earnings of low-income native workers than if it was motivated by racial animus. To our knowledge, economists have not yet studied the effects of policymakers' perceived motivations on how observers judge the fairness of their policies.

Finally, our analysis relates to ongoing debates among both economists and legal scholars about which forms of discrimination are more 'egregious' than others (and therefore perhaps more deserving of policy remedies or legal sanctions.) For example, in a recent review article, Bertrand and Duflo (2017) provide the following description of a common view among economists:

> While taste-based discrimination is clearly inefficient…, statistical discrimination is theoretically efficient and, hence, more easily defendable in ethical terms under the utilitarian argument. Moreover, statistical discrimination can also be argued to be "fair" in that it treats identical people with the same expected productivity (even if not with the same actual productivity) [equally] and is not motivated by animus. In fact, many economists would most likely support allowing statistical discrimination as a good policy, even where it is now illegal… (Bertrand and Duflo 2017, p. 312).[4]

In the case of legal debates and proceedings, influential decisions like Griggs v. Duke Power Co. (1971) have maintained that *intent* is not essential for an act or policy with race-based consequences to be unlawful, instead these decisions maintain that disparate *impact* is enough. This disparate impact principle continues to be contested, however.[5] Our paper contributes to both these economic and legal debates by describing how *a broad sample of Americans* perceives the fairness of different types of discriminatory actions. We find that (a) the detailed intentions underlying a discriminatory action *do* matter for peoples' fairness perceptions, but that (b) whether the action was motivated by someone's racial animus ('tastes') is not, on its own, a reliable guide to an action's perceived fairness.

Section 2 of the paper describes our survey design, data collection, and sample characteristics. Section 3 presents some basic facts about fairness perceptions: How do perceptions vary with respondent characteristics, survey treatments, and interactions between the two? Section 4 describes four simple, preregistered models of fairness and compares their implications to subjects' aggregate response

[4] The word "equally" is not present in Bertrand and Duflo's text; we have inserted it to convey what we believe is their meaning.

[5] Despite these disputes, there seems to be wide agreement that the presence of racial or other animus would make the same discriminatory act *more* egregious.

patterns. It shows, --among other things—that our Group 2 respondents (*Utilitarian*s) care about two criteria –utilitarianism and race-blindness-- that sometimes conflict. Section 5 then uses within-subject variation in the discriminatee's race to estimate the relative weight these *Utilitarian* subjects place on the two criteria they care about. Section 6 concludes.

2. Design and Implementation

2.1 Survey Structure

Before starting our survey, all our subjects were informed that they will be exposed to four scenarios, with the proviso that "Some of these scenarios may seem realistic to you; others may seem unrealistic." We also told subjects that only very limited information about each scenario will be provided. Nevertheless, subjects were asked to "please give us your reaction to [the scenarios] if they were to happen, based on the information that has been provided". The goal of these statements was to clarify that we want respondents to assess the *fairness* of the hypothetical interactions (and not their realism or their likelihood of occurring).

Next, our subjects are randomly assigned to read four vignettes in which an employer, "Michael" (or "Andrew") makes a hiring decision between a White and a Black applicant.[6] These scenarios are designed to represent canonical examples of taste-based and statistical discrimination. We define taste-based discrimination as a decision that is based on *someone's* racial animus, and distinguish two forms: *Less-justifiable* taste-based discrimination is based on the *employer's own distaste* for people of a particular race. *More-justifiable* taste-based discrimination occurs when the employer accommodates his *customers'* distastes for a particular race.[7] Statistical discrimination, on the other hand, is based on differences in expected job performance. *Less-justifiable* statistical discrimination is based on low-quality information about the relative performance of two racial groups; we frame this as a non-

[6] Michael and Andrew appear to be the most common male names that are relatively race-neutral. Between 2011 and 2016, they ranked in the top 2-6 names for White men and the top 6-12 names for Black men in New York City birth names.

[7] Notice that –to the extent that it is costly to attract new customers-- our employer's decision is profit-maximizing in the more-justifiable version of taste-based discrimination, but not in the less-justifiable version.

quantitative statement from a single, non-expert source (a 'neighbor') about problems *others* experienced when employing White or Black employees, such as lateness and lack of attention to detail. *More-justifiable* statistical discrimination is based on "reliable statistics" from "a large and experienced network of local business owners" who frequently hire for the same type of opening as in the scenario.[8] In both cases, the 'justifiability' rankings of the more detailed reasons for the action are based on our own priors regarding how respondents would react. After reading each vignette, the respondent is asked to rate the fairness of the employer's hiring decision on a scale from 1 to 7. As in Alesina et al. (2020, 2021), these questions have no material consequences.[9]

The four scenarios encountered by each respondent are presented in two Stages. In Stage 1, subjects were assigned with equal probability to one of the four possible treatment combinations: SW, TW, SB, and TB, where S and T represent statistical and taste-based discrimination, and W and B indicate the race of the discriminatee. Within Stage 1, the subjects encounter the less- and more-justifiable versions of discrimination in random order. In Stage 2, subjects were randomly assigned to one of the three treatment combinations they did not encounter in Stage 1, and again encountered the more- versus less-justifiable forms in random order.[10] Thus, as illustrated in Appendix A1.2, two thirds of the respondents encountered a switch in the discriminatee's race, and two thirds encountered a switch between taste-based and statistical discrimination.

Our survey concludes with Stage 3, which asks all subjects the same questions. First, in an open-text question, we remind respondents of the final scenario they encountered and invite them to explain the

[8] As framed in our vignettes, low-quality information could be interpreted by our subjects as an unbiased signal with high variance, as a biased signal (Bohren et al. 2019), or even as a signal whose bias is motivated by someone's racial animus. In all cases we would expect that relying on such signals will be seen as less fair than using the information described in our high-quality statistical scenario.

[9] Cappelen et al. (2019) and Almås et al. (2020) create real *income* inequality (for example, between two MTurk workers), then give third-party subjects options to reduce this inequality. Applying this methodology to discriminatory incidents would raise serious ethical concerns. It is also unclear how we could manipulate the *motives* of a real discriminator, which matter a lot for peoples' fairness assessments.

[10] To make the scenarios more realistic, the name of the employer also switches between the Stages. Specifically, half the employers are Michael and the other half Andrew in Stage 1; this assignment is random. In Stage 2, the name of the employer switches to the other, unused name for all respondents.

fairness assessment they made. Next, we use the following question to elicit subjects' assessment of Black people's relative economic opportunities (BRO):

> 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? Black people have:*
>
> *Much less / Less / A Little Less / Roughly equal / A little more / More / Much More opportunity than White people.*

Finally, we collected information on the subjects' age, education, race, gender, and political affiliation.

2.2 Scenarios and Fairness Assessments

To illustrate how our fairness assessments work, we next describe Stage 1 of the survey for subjects who are assigned to the TB (Taste, Black) treatment combination. To introduce this Stage, we first tell subjects they will encounter two scenarios which share many common elements but contain some differences; we also say that the differences have been underscored to make them easier to pick out. The subjects then read and assess the *less* or *more* justifiable forms of the Taste discrimination scenario with a Black discriminatee in random order. The *less* justifiable form of taste discrimination is motivated by the employer's 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.

The *more* justifiable form is identical, except for the following underscored sections:

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.

Michael decides to hire the White worker, in order to avoid losing sales to customers who do not want to interact with Black representatives.

After each scenario, the respondent is asked to "indicate the extent to which you thought that Michael's hiring decision was fair" on a seven-point scale, where 1 was "very unfair", 4 was "neither fair nor unfair", and 7 was "very fair".

As noted, in Stage 2, respondents encountered two more scenarios in which either the *race* of the discriminatee (Black or White), the *motivation* for the discrimination (Tastes versus Statistical) or both of these were different from Stage 1.[11] White scenarios were identical to Black scenarios except that the races of the discriminator and the discriminatee are reversed. As noted, *less* justifiable statistical discrimination was based on low-quality information (hearsay from a single, uninformed source) about relative group productivity, and *more* justifiable statistical discrimination was based on higher-quality information (quantitative information from substantial sample of other employers). The exact wording of these and all our scenarios is provided in Appendix 1.1.

[11] Exposing subjects to four scenarios (rather than one) has three main benefits. First, it gives us more fairness assessments, while preserving the option to use only each subject's first treatment for pure cross-subject comparisons. Second, as illustrated in Section 5.1, it allows us to test for a specific but plausible form of experimenter demand effects. Third, as illustrated in Section 5.3, it allows us to assess the relative importance subjects assign to utilitarian, versus race-blind fairness criteria when those criteria conflict.

2.3 Implementation and Representativeness

On September 21, 2020, we pre-registered our survey design and procedures, and posted a pre-analysis plan. Our survey was administered to a sample of MTurk workers between September 22 and October 6, 2020. Subjects were given one hour to complete the survey and were informed that we expected the task to take about 15 minutes. Conditional on completing the entire survey, subjects were paid $5.[12] A few measures were taken to improve the accuracy and representativeness of the responses. First, respondents were required to have a U.S. address. Second, to further discourage foreign workers from participating, the survey was launched during U.S. Pacific daylight hours on weekdays. Third, MTurk workers were required to have a 95 percent approval rating to discourage robots (i.e., automated responses). Fourth, the survey included a CAPTCHA question to further discourage robots. Finally, respondents were exposed to each vignette for at least 30 seconds before being allowed to submit their fairness assessment. In all, we received 779 responses; during data cleaning we dropped 137 of these, leaving us with a final count of 642 responses in our analysis sample.[13]

In Appendix 2.2, we present summary statistics of our survey respondents and compare them to adults in the 2019 American Community Survey (ACS) and the 2020 General Social Survey. Compared to the ACS, our sample of MTurkers is quite regionally representative, a little more male, and a little more likely to be either White or Black. Our respondents are also considerably better educated and much more likely to be between 25 and 44 years of age than U.S. adults in general. For the most part, these are well known features of the MTurk population.[14] Comparing our subjects' political orientations to the GSS is more difficult because –despite the similarity of the survey questions—the middle category differs

[12] In comparison, the average effective hourly rate on MTurk is about $4.80 (Kuziemko et al., 2015). The average actual survey completion time for our subjects was 11.5 minutes.

[13] The main reasons for excluding responses were (i) a pinged location suggesting that the respondent was not U.S. based, and (ii) indications that the response was automated (for example, the IP address attempted our survey more than once, or the response copied and pasted word-for-word sentences from the vignettes into their open text answer.) Summary statistics on the final sample's characteristics (race, gender, education, political orientation, and location within the U.S.) can be found in Appendix 2.

[14] For additional discussions of the representativeness of MTurk samples, see Kuziemko et al. (2015), Arechar et al. (2017), and Everett et al. (2021).

between the two surveys: "moderate" in our case versus "moderate, middle of the road" in the GSS.[15] Ignoring this difference in phrasing, it would appear that our MTurk respondents are politically more 'extreme' than GSS respondents, with more candidates selecting the two extreme categories and far fewer selecting the middle one. However, GSS respondents could be attracted to the attracted to the 'middle of the road' label.[16] In sum, our MTurk-based sample differs from the U.S. population in substantial and mostly well-known ways. In Appendices 7 and 8 we estimate the implications of these differences by re-weighting our main results to match the American Community Survey and General Social Survey respectively. The results are very similar.

2.4 Question Order Effects

In all multi-part surveys, but especially in contexts like ours where framing and experimenter demand effects might play a large role, the order in which respondents encounter different questions could have large effects on the respondents' answers. We address this issue in detail in Appendix 3, which shows that question order effects are absent from our survey in two distinct senses. First, as shown in Appendix 3.1, there is no time trend in fairness assessments across the four scenarios encountered by each respondent: Respondents become neither more nor less accepting of discrimination as they are asked additional questions about it.[17] Second, the order in which respondents encounter the Tastes versus Statistical and the *more v*ersus *less justified* scenarios does not affect their fairness ratings on subsequent scenarios. In Appendices A3.2 and A3.3, this is illustrated three different ways: First, we show that subjects' subsequent assessments of a given type of discrimination (e.g., Taste) do not depend on which

[15] 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 political representativeness of our population. Our political party preference question is not comparable to the GSS's, but (with the exception of this middle category) our political leaning question is identical to the GSS's (see Table A2.2 for details).

[16] In addition to this possible difference in variance, there is also some suggestion that, on average, MTurkers are somewhat more liberal than GSS respondents. 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). This difference could reflect the relative youth of MTurkers.

[17] Recall that treatments are assigned in a balanced way across the four scenarios each respondent encounters, so aggregate comparisons of fairness ratings over time are not contaminated by changes in the mix of scenarios people encounter.

type (Tastes or Statistical) they encountered previously. Second, we cannot reject that the fairness ratings *changes* of subjects who switched from, say, a *more to a less justified* treatment were equal but opposite in sign to subjects who switched in the opposite direction. Finally, for both the type of discrimination and the *justifiability* treatments we show that within-subject, between-subject and pooled fairness regression estimates are statistically indistinguishable from each other.[18]

The one treatment that does, however, affect subjects' subsequent fairness assessments is the *race* of the discriminatee. As we document in Appendix A3.4, our respondents' Stage 2 fairness assessments depend on the *race* treatment they encountered in Stage 1. In the next two Sections of the paper (3 and 4), we will eliminate the influence of these order effects by relying only on data from Stage 1 of the survey: There is no within-subject variation in the *race* treatment during Stage 1 (or during Stage 2), because discriminatee race only varies between the experiment's two Stages. In Section 5, we will document and scrutinize these order effects in greater detail, and exploit them to shed light on the tension between two models of fairness (utilitarian versus race-blind rules) among the moderate and liberal respondents to our survey.

3. Some Facts

This Section describes how fairness perceptions in our survey depend on the respondent's personal characteristics, on the experimental treatment the respondent encountered, and on some interactions between these (for example, between the respondent's political orientation and the race of the fictitious discriminatee). As already noted, to avoid any influence of treatment order effects for the *race* treatment, this entire Section uses only responses from Stage 1 of the survey, giving us two responses per subject.[19] To account for within-subject correlation of responses, all standard errors are clustered by subject.

[18] 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 treatment each respondent encountered only. Pooled estimates include all four scenarios each person encountered, without person fixed effects.

[19] There is no within-subject variation in the race treatment within Stage 1 (or within Stage 2). All of the comparisons described in this Section continue to apply if we go even further and use only data from the very first of

3.1. How Does Perceived Fairness Vary with Respondents' Characteristics?

In Figure 1, we show how the mean perceived fairness of discriminatory acts varies with respondents' characteristics. To maximize our sample size for these initial comparisons, we pool responses across all four treatment combinations (SW, TW, SB, and TB) as well as the more- and less-justifiable versions of both types of discrimination. In short, Figure 1 shows that our subjects' mean fairness assessments do not vary significantly with their age or race. However, women viewed the discriminatory acts as slightly less fair than men. Somewhat surprisingly (to us), respondents' fairness assessments were positively related to their education levels; we explore this correlation in Appendix 4 and show that higher levels of education mostly reflect a higher 'set point' for all fairness assessments in the following senses. First, regardless of political leaning, more-educated individuals rate *all* the scenarios they encounter as more fair than less-educated individuals. Furthermore, in contrast to political leaning –which has strong effects on how our subjects respond to some of our treatments-- highly-educated subjects respond to all our treatments in very similar ways to less-educated subjects.[20]

Finally, Figure 1 shows that respondents' political leaning is strongly related to the perceived fairness of discriminatory acts. Self-described conservative respondents perceive these actions to be fairer than both moderates and liberals (e.g., $p = .000$ for conservatives versus liberals).[21] Mean fairness assessments across U.S. political *party* preferences (e.g., Democrats versus Republicans) exhibit similar patterns.[22] Since the Independent group could include people with both extreme right- and left-wing

---

the four scenarios each person encountered, although the standards errors are somewhat higher. See for example Figure A5.1.1, which replicates Figure 1 using first-scenario data only.

[20] For example, Appendix 4 shows that more-educated respondents' greater tolerance of discriminatory acts is not confined to discrimination against a particular race—it applies equally to anti-Black and anti-White discrimination. We also show that educated peoples' higher fairness assessments are not related to differences in political affiliation across education categories: The positive association between education and overall fairness ratings remains very strong within both conservative and liberal survey respondents.

[21] To account for the fact that our data contain multiple observations per respondent, all p-values in the paper are clustered by respondent.

[22] There is a statistically insignificant non-monotonicity with respect to party preference, with Independents being more opposed to discrimination than Democrats.

orientations, we use conservative-liberal leaning rather than party affiliation to categorize respondents' political preferences in the remainder of the paper.[23]

3.2 Treatment Effects

In Figure 2, we compare the fairness assessments of subjects who were exposed to the Tastes versus Statistical treatments, and to the more versus less justifiable forms of each. As in Section 3.1. we pool both of the Stage 1 scenarios encountered by each worker and cluster our standard errors by respondent. To simplify the presentation, we also pool the Black and White treatments.[24] To facilitate interpretation here and throughout the paper, we report all fairness assessments on a scale from -3 ("very unfair") to 3 ("very fair"), where 0 was labeled in the survey as "neither fair nor unfair."[25] The standard deviation of fairness assessments in Stage 1 is 1.657 across respondents, 0.961 within respondents, and 1.915 overall.

According to Figure 2, the average respondent sees no meaningful distinction between the fairness of the statistical versus taste-based scenarios in our survey ($p = .971$). Conditioning on whether discrimination is taste-based or statistical, however, subjects view the less justifiable form of either taste-based or statistical discrimination as less fair than the more justifiable form ($p = .000$ in both cases), confirming our expectations. To illustrate the size of these differentials, we first remark that an average respondent did not view the more-justifiable forms of either statistical or taste-based discrimination (high quality information; accommodating the tastes of others) as unfair at all: the mean fairness ratings of these actions were in the "somewhat fair" range with small standard errors.[26] In contrast, the less

[23] Of the respondents who identify themselves as "Independent" within our sample, about 8.43% suggest they are either "extremely liberal" or "extremely conservative." Furthermore, all the results by political party are very similar, with occasional non-monotonicities similar to Figure 1(e), where Independents appear to be to the left of Democrats.

[24] Figure 3 shows that the effects of *justifiability* are virtually identical for White versus Black discriminatees.

[25] As noted, the subjects saw these verbal descriptions, associated with the numerals 1 through 7.

[26] The confidence interval for the fairness of *more*-justifiable taste-based discrimination includes zero (neither fair nor unfair); for *more*-justifiable statistical discrimination the confidence interval is bounded above zero.

justifiable forms of taste and statistical discrimination were both viewed much more harshly—specifically 0.925 units (on a scale of -3 to 3), or 0.483 standard deviations less fair.

In Figure 3 we turn our attention to the *race* treatment—i.e. the race of the person who was discriminated against. Motivated by Figure 2 (which shows no difference between the Statistical and Tastes treatments) we now pool these treatments but continue to distinguish between their more- versus less-justifiable forms. In the sample as a whole, Figure 3 shows that respondents view the same discriminatory acts more negatively when they are directed at Black than at White job applicants. We call this phenomenon the *Discriminatee Race Effect* (DRE). The DRE shown in Figure 3 is substantial in magnitude, amounting to about 0.5 fairness units or 0.263 standard deviations, and highly statistically significant (p= .002 and .000 within the *less* versus *more* justifiable forms of discrimination, respectively).

3.3 Heterogeneity: Discriminatee Race Effects by Respondent Race and Political Orientation

While the effects of the *race* treatment shown in Figure 3 are interesting, these effects may not be the same for all types of respondents. For example, Black respondents might react more negatively than White respondents to discrimination against Black job applicants. To explore this issue, Figure 4 presents separate estimates of the discriminatee race effect for respondents of different races. Unfortunately, our samples of both Black and Other racial groups are too small to precisely estimate a discriminatee race effect within either group. The point estimates for these groups however suggest that both groups respond to the race of the discriminatee in much the same way as White respondents do.[27] In sum, Figure 4 underscores the fact that the discriminatee race effect in our data – i.e., the tendency to see discrimination against Black people as less acceptable than discrimination against White people—is driven primarily by our White respondents, who comprise about 78 percent of the sample. Thus, while we continue to estimate all our results on the full sample of MTurk respondents in the remainder of the

[27] Interestingly, Figure 4 indicates that the Other group views discrimination relatively harshly. However, there is little indication of a discriminatee race effect for this group of respondents (p = .506) and the point estimates themselves are imprecise.

paper, it is important to bear in mind that the stark political differences we will document throughout the paper are driven, to a substantial degree, by differences between White respondents with different political leanings.

Turning to those political differences, Figure 5 presents separate estimates of the discriminatee race effect by the respondent's political leaning. These reveal a clear difference: the discriminatee race effect is stronger among moderate and liberal respondents than in the sample as a whole, but is absent among conservatives. Conservatives view discrimination against (fictitious and identically qualified) Black and White job applicants the same way: as more fair than unfair.[28] A final striking finding from Figure 5 is the strong similarity in both the levels of fairness rankings and in the discriminatee race effects between self-described moderate and liberal respondents. Later in the paper (starting in Section 4.4) we exploit this fact to simplify our analysis by comparing just two political groups—conservatives versus moderates/liberals.

4. Assessing Four Models of Fairness

This Section describes four simple models of how subjects might evaluate the fairness of discriminatory actions: (simple) utilitarianism, racial in-group bias, race-blind rules (RBR), and belief-based utilitarianism (BBU). For each model, we compare its predictions with the main empirical patterns in our data and show that two of the models –racial in-group bias and BBU-- are inconsistent with some key patterns in our data. After examining subjects' open-text responses for clues that might explain these inconsistencies, we then propose a two-group framework with three fairness criteria that does a better job of accounting for the facts we have documented. As in Section 3, our analysis only uses data from Stage 1 of the experiment to ensure that *race* treatment order effects cannot affect our conclusions.

[28] The confidence interval for anti-Black discrimination is bounded above zero; the mean fairness assigned to anti-White discrimination is almost identical, but not quite significantly different from zero.

4.1 (Simple) Utilitarianism

In general, utilitarian models of fairness share two main features, the first of which is that fairness depends on outcomes, not on intentions or justifications. Since our Tastes vs. Statistical and less-versus-more justifiable treatments refer to the *reasons* for the employer's actions, and since the consequences of the employer's actions –i.e. which worker got the job—are statistically balanced across all our observations, the fairness ratings of purely utilitarian respondents should not differ between any of the Taste-based, Statistical, or less- and more-justifiable forms of discrimination. Second, utilitarian fairness models use a social welfare function to map consequences into fairness levels. In the simple utilitarian model we consider here, this social welfare function is a strictly concave function of the incomes of the people depicted in our scenarios. Since mean racial income differences which indisputably favor White people, utilitarian respondents should be more tolerant of anti-White than anti-Black discrimination. Importantly, this prediction holds even if we account for the direct effects of discrimination on employers' utility: Taste-based discrimination may raise the utility of the employer, and statistical discrimination may raise profits. This because the employer in our scenarios is always White when the discriminatee is Black, and *vice versa*.

We refer to the type of utilitarianism described in this subsection as 'simple' because it is based purely on racial *income* differences, which are publicly verifiable information. Real respondents might, however, base their ideas of deservingness on criteria other than income –such as *opportunities*—and could have inaccurate and widely varying beliefs about racial gaps in income and opportunity. (Davidai and Walker 2021, Kraus et al. 2017, 2019). We will consider these possibilities under the heading of *belief-based utilitarianism* (BBU) below.

Turning to the evidence on simple utilitarianism, Figure 3 has already shown that respondents in general *do* view discrimination against Black applicants more harshly than discrimination against White applicants. That said, Figure 5 showed that this tendency was confined to moderates and liberals: Conservatives do not consider race when assessing the fairness of discriminatory actions. We conclude that our (simple) utilitarian model is consistent with moderates' and liberals' fairness assessments, but is

not consistent with conservatives' fairness statements.[29] Utilitarianism also cannot account for the large justifiability effects on perceived fairness that are documented in Figures 2 and 3.

4.2 Racial in-group bias

The phenomenon of in-group favoritism, where people value actions that benefit members of their own identity group more than actions benefiting others, has been extensively documented (Luttmer 2001, Chen and Li 2009, Fong and Luttmer 2009, 2011, Everett et al. 2015). While a variety of models could explain this behavior, a simple one, based on social preferences, modifies the preceding utilitarian model in a straightforward way: instead of favoring actions that benefit lower-income groups, persons motivated by racial in-group bias will favor actions that redistribute resources from members of other races to members of their own. In our experiment, respondents who exhibit racial in-group bias should view the discriminatory acts we depict as less fair when the fictitious discriminatee shares the respondent's race.[30]

As Figure 4 has already shown, we do not have the statistical power to test these predictions for the respondents in our Black or Other racial categories.[31] Our evidence for White respondents, however, is strongly inconsistent with racial in-group bias: As a group, White respondents view discrimination against Black people as substantially *less* fair than discrimination against White people. Interestingly, when we focus our attention on the subset of White respondents who identity as conservative, this strong rejection of in-group bias no longer holds: As shown in Figure A5.2.1, White conservatives rate discrimination against Black people as 0.405 units *more* fair than discrimination against White people.

---

[29] An insignificant discriminatee race effect for conservatives could imply that that they are not utilitarians at all (i.e. they do not use a social welfare function (SWF) to make fairness assessments). Alternatively, conservatives could be utilitarians with a linear SWF. A final possibility, considered under *belief-based utilitarianism* (BBU) below, is that conservatives' SWF depends on something other than income (such as, for example, perceived relative opportunities).

[30] Related (and with the same empirical predictions in the case of our experiment) we would also expect respondents to more forgiving of discriminatory acts committed by a member of their own racial group.

[31] In this respect, our MTurk sample is no different from any nationally representative sample of this size. Without quota-sampling minority respondents (which is not possible on MTurk) a much larger sample would be needed to measure the amount of in-group racial bias among other racial groups.

This difference is however not statistically significant at conventional levels ($p = .134$). Overall, we conclude that racial in-group bias model does not provide a useful lens for understanding the main fairness ratings patterns we have documented.

4.3 Race-Blind Rules (RBR)

In contrast to utilitarianism and in-group bias, *rules-based* models of fairness are not consequentialist in nature; instead, they belong to the class of *deontological ethics*, which associate fairness with adherence to a consistent set of rules (Andreoni et al. 2019). Further, in deontological ethics, *intentions* can matter and consequences are secondary: for example, ill-intentioned actions that unintentionally produce a good outcome are considered unethical. Intent and motivation play key roles in civil and criminal law, and abundant evidence from behavioral economics shows that people care about intentions when assessing the fairness of many economic actions.[32] Finally, rules-based models of fairness are *race-blind* when the rules that assign fairness to actions and intentions do not depend on the races of the people involved.

Applying these ideas to our experiment, an RBR model of fairness would – unlike the previous two models – allow the fairness of a discriminatory action to depend on the intentions behind it: Did the act serve to indulge the employer's personal racial animus, or to protect his business from retaliation by racist customers? Did the employer do his due diligence before relying on statistical information in hiring, or did he a take hearsay-based shortcut? Further, assuming the respondent has an implicit set of rules defining which of the above motivations are fairer than others, she should apply those rules in a race-blind way. For example, if using low-quality statistical information is x units less fair than using high-quality information, x should be the same regardless of the race of the discriminatee.

[32] Intentions are relevant to the distinction between first- and second-degree murder, for example. Charness and Levine (2000), Offerman (2002), Abeler et al. (2010) and Breza et al. (2017) document the effects of intentions on peoples' reactions to layoffs, pay reductions, and pay inequality.

The fairness ratings of our respondents are consistent with the use of race-blind rules (RBRs) in at least three ways.[33] First, the effects of our *justifiability* treatments in Figure 2 strongly support the idea that respondents care about the employer's motivation for discriminating against a job applicant. Importantly, our experimental design ensures that the *justifiability* effects in Figure 2 hold the consequences of the discriminatory action constant: While the material consequences of not being hired could, for example, vary with the discriminatee's race (because of differences in outside labor market options), notice that Figure 2 varies only the *reasons* for not being hired: discriminatee race is balanced between the motivation and justifiability treatments due to random assignment.

Second and more strikingly, Figure 3 shows that our respondents penalized the less-justifiable forms of discrimination by the same amount (relative to the more justifiable forms), *regardless of the race of the discriminatee*: (-0.953 versus -0.898 fairness points for White versus Black discriminatees respectively, with $p = .679$ for a test of equality). Third, a similar test shows that this stability to discriminatee race also applies to the Taste/Statistical fairness differential—it is essentially zero for both Black and White discriminatees.[34] A final, remarkable feature of our respondents' apparent adherence to race-blind rules is that it applies just as strongly on both sides of the U.S. political divide. We show this explicitly in Figure 6, which shows that respondents ranked the relative fairness of *more* versus *less* justifiable forms of discrimination almost identically, irrespective of their political leaning.

In sum, there is substantial *prima facie* evidence of deontological ethics based on race-blind rules among our subjects: Subjects care about the reasons why a discriminatory act occurred in a consistent manner (Tastes versus Statistics *per se* do not matter; other motivational factors captured by our *justifiability* treatments do matter). Consistent with a widely held desire to adhere to race-blind rules, these motivational factors affect the perceived fairness of a discriminatory action in strikingly similar

[33] In Sections 5.2 and 5.3, we will present a third piece of evidence supporting the RBR model that applies only to moderate and liberal respondents. Specifically, we will argue that the order effects for the Black treatment (which are present only for moderate and liberal respondents) suggest that these respondents prefer to maintain a form of consistency across race in their fairness assessments.

[34] Within Black Discriminatees, Tastes-Based scenarios are 0.121 units more fair. Within White Discriminatees, Taste-Based scenarios are 0.138 units less fair. A test for equality of the Tastes vs. Statistical gap between the Black and White treatment yields $p = .319$.

ways regardless of the race of the discriminatee, and regardless of the political orientation of the respondent.

4.4 Belief-Based Utilitarianism (BBU)

In Section 4.1 we ruled out (simple) utilitarianism among conservative respondents because those respondents did not object more strongly to anti-Black than to anti-White discrimination, even though Black job applicants, on average, have lower incomes. This fact, however, does not rule out the possibility that conservatives are motivated by a different form of utilitarianism, which we label *belief-based utilitarianism* (BBU).[35] Under BBU, respondents still value redistribution from more- to less-advantaged groups, but they use a different and possibly subjective metric of relative advantage to guide their fairness evaluations.[36] From a modeling perspective, BBU is an appealing hypothesis because it would allow us to explain a key empirical difference between conservatives and other respondents—conservatives do not exhibit a discriminatee-race effect—in a straightforward way: Both conservatives and other respondents are in fact utilitarians (i.e. they prefer to favor a disadvantaged group) but they simply have different beliefs about who is disadvantaged.

Evidence that is consistent with BBU is presented in Figure 7, which draws on the BRO question in Stage 3 of our survey. This question asked the respondents to rate Black people's relative economic opportunities in the United States on a seven-point scale, running from "much less opportunity" (minus 3 in Figure 7) to "much more opportunity" (plus 3 in Figure 7). Figure 7 shows that the respondents' BRO ratings differ dramatically by their political orientation: While liberals have a mean BRO of -1.374 (p = .000), conservatives' mean of -0.206 is insignificantly different from zero (p = .089) with moderates in between. This suggests that conservatives' belief that Black and White people have roughly equal

[35] BBU is essentially the conceptual framework laid out in Alesina et al. (2020), and underlying the empirical work in Alesina et al. (2021): People have beliefs about the relative incomes and opportunities available to different demographic groups, then use a utilitarian ethic (favoring the lower-opportunity group) to translate these beliefs into support (or non-support) for public policies.

[36] Our survey design does not allow us to distinguish whether respondents' beliefs about relative opportunities motivate their perceptions of the fairness of discriminatory acts, or whether these beliefs are *motivated by* a desire to evaluate discriminatory actions in a certain way. Oprea and Yuksel (2021) use a cleverly designed experiment to detect motivated beliefs in a different context from ours.

opportunities has the potential to account for their observed fairness ratings, which –like their BRO ratings—are statistically the same for discrimination against Black versus White job applicants.[37]

To assess whether BRO differences can actually account for the partisan gap in fairness assessments, panel (a) of Figure 8 shows respondents' fairness ratings for anti-Black discrimination by BRO categories, separately for conservatives and moderates/liberals.[38] If BRO accounts for the large partisan gap, we should see little or no partisan gap *within* the BRO categories: Instead, the partisan gap should be explained by the higher mean level of BRO among conservatives. The evidence, however, paints a very different picture in two key respects. First, while BRO is very predictive of the perceived fairness of anti-Black discrimination among *moderates and liberals*, it is not predictive of conservatives' fairness ratings. In other words, we see no effect of BRO on perceived fairness of anti-Black discrimination among conservatives, even though their beliefs about relative racial opportunities vary widely. Second, Figure 8 shows that there are large political gaps in the perceived fairness of discriminating against Black people, even when we condition on BRO. These political gaps are particularly stark at the bottom of the BRO distribution: While moderates and liberals who think that "Black people have much less opportunity than White people" (BRO=-3) are strongly opposed to anti-Black discrimination, conservatives with the same belief are, on average, *accepting* of anti-Black discrimination (with a mean fairness rating of about +0.5). This partisan gap at the bottom of the BRO distribution is highly statistically significant. Within subjects who have BRO levels of -3, and within subjects who have BRO levels of -2, the partisan gap is significant at p=.000.

---

[37] Our findings about the partisan gap in perceived relative opportunities (BRO) mirror the partisan differences in perceptions about inequality and mobility documented by Alesina et al. (2020), and the stark partisan differences in beliefs about the causes of racial inequality documented by Alesina et al. (2021). They also mirror Alesina et al.'s (2021) and Haaland and Roth's (2021) findings that Democrats perceive that there is much more anti-Black discrimination than Republicans do. As noted, our contributions relative to these papers are that we study the fairness of individual (discriminatory) actions (not public policies), we demonstrate the key role of the intentions behind an action in determining its perceived fairness, and we test the BBU model that underlies the idea that changing beliefs about opportunities can change support for policies.

[38] Starting in this subsection, we combine moderates and liberals into a single group to simplify the presentation and preserve statistical power. Interested readers can view a version of Figure 8 with all political groups in Appendix 5.3; all the qualitative results discussed below are similar for moderates and liberals individually, as well as the combined group.

A third and even more surprising piece of evidence against the "BRO hypothesis" emerges from panel (b) of Figure 8, which replicates panel (a) for discrimination against White job applicants. Consistent with a large explanatory role for BRO in peoples' fairness assessments, this Figure shows an effect of BRO on the perceived fairness of discrimination that is essentially invariant to political orientation: the coefficients are .257 (p = .094) and .265 (p=.000) for conservatives and moderates/liberals respectively. However, for both political groups the direction of this effect is the opposite of what the BRO hypothesis would predict: According to the BRO hypothesis, higher levels of Black people's perceived relative opportunities should make discrimination against White people less acceptable. Instead, the perception that Black people have equal or more economic opportunities than White people – which is held by 36.9 percent of our subjects—is associated with a *greater* tolerance of (hypothetical) acts of anti-White discrimination.

Summing up, while respondents' stated beliefs about Black peoples' relative opportunities (BRO) are (a) correlated with their political affiliations and (b) sometimes predictive of their fairness assessments, the signs and patterns of these associations are decidedly inconsistent with the 'BRO hypothesis': the idea that conservatives' beliefs about Black relative opportunities explain their tolerance of anti-Black discrimination. This rejection of the BRO hypothesis in our data might help explain why interventions designed to change beliefs about relative opportunities do not have robust effects on support for race-based policies, *even when the interventions change beliefs* (Alesina et al. 2021; Haaland and Roth 2021).

4.5 What Motivates Respondents Who Say Discrimination is Fair?

To try to make sense of the unexpected findings in Figure 8, we first observe that a large subset of our respondents (when classified by beliefs and political orientation) exhibit a common pattern of fairness assessments: they assign roughly equal, non-negative fairness to both anti-Black *and* anti-White discriminatory acts. In Figure 8, this group of respondents includes all self-identified conservatives, *plus* the moderates and liberals with BRO ≥ 0. Together, these respondents (henceforth Group 1) represent 48.9 percent of all respondents. In Figure 8 they are indicated by the hollow circle markers.

To understand what types of fairness criteria might account for these respondents' fairness assessments, we then turned to our respondents' open-text explanations of the last scenario they encountered. As described in Appendix 6, we manually classified these explanations into three broad categories, separately for respondents who rated discrimination as "unfair" or "very unfair", versus respondents who said it was "fair" or "very fair". For the latter group, the most common type of response was some variation of '*the business must thrive*', such as:

> "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."
>
> "A business wants [to] retain customers and high profits. So give them what they expect. Hiring what people prefer is reasonable."
>
> "Andrew needs to do what is best for his business. If he hired the black worker, he'd lose money and perhaps even go out of business."
>
> "To ensure the success of his business, Michael should do everything possible to do so. Is it racist? I don't think so. Michael is free to hire whoever he wants."

Notably, almost all of these 'business must thrive' answers referred to scenarios in which the employer accommodated his customers' discriminatory tastes (i.e. the 'more-justifiable' version of taste-based discrimination).

Closely related, the second most common type of explanation involved some statement of '*employer rights*', including:

> "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."
>
> "It's his business. He doesn't need to justify to me any of his hiring practices. … Seriously, he does not need to justify his hiring choices."
>
> "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."
>
> "I've never had a problem with this as long as every business owner is allowed to do it; I don't feel comfortable in all-black establishments so just I don't go to them. They'd be uncomfortable, I'd be uncomfortable, just let them have their thing. What's the problem?"
>
> "Employers are allowed to hire whoever they see as best for the job."

Notably, these 'employer rights' explanations were expressed with respect to *all* forms of discrimination, including discrimination based on the employer's own tastes.

Taking all the above responses together, we propose that Group 1's high acceptance of discriminatory actions – regardless of the target of the action – is consistent with a fairness system that prioritizes individual decision rights (regardless of those decisions' effects on others), especially for business owners. As shorthand, we therefore label Group 1, defined above, as *Business Rights Advocates.*

4.6 An Interpretive Framework that 'Works'

With this decision-rights-based ethical model in hand, we can now propose a provisional, ex post interpretive framework that ties together all the fairness assessment patterns we have documented in the paper. We hasten to remind readers that –by logical necessity—this framework is not the only one that can account for all those patterns. We offer it primarily as a simple mnemonic device that summarizes the main facts we have established, and as a jumping-off point for future research on these questions.

In our proposed framework, there are two main groups of respondents. Group 1 (the "Business Rights Advocates", accounting for 48.9% of all respondents) includes all conservatives**,** plus moderates and liberals with BRO ≥ 0. These respondents are, on average, accepting of all the discriminatory actions we depict, regardless of the race of the discriminatee. Members of Group 1 justify these fairness assessments as protecting or raising profits and preserving employer rights. Members of Group 1 do not appear to be motivated by any utilitarian concerns (either 'simple' or 'belief-based'). Based on our

findings in Section 4.3, however, they share the widespread support across the political spectrum for race-blind procedural fairness (RBRs).

Group 2 or "Utilitarians" are moderates and liberals with BRO<0, accounting for 51.1% of our respondents. These respondents object to both anti-Black and anti-White discrimination, but object more strongly to anti-Black discrimination. Given their beliefs (BRO<0), Group 1's fairness assessments are consistent with both simple and belief-based utilitarianism. This group exhibits no obvious support for employer decision rights. [39] Like Group 1, Group 2 shares a strong desire for race-blind procedural fairness.

5. Reconciling Conflicting Fairness Criteria: Utilitarianism versus RBRs

In the previous Section, we proposed a two-group interpretive framework in which one group of respondents (Group 1 – *Business Rights Advocates*) cares only about race-blind rules (RBRs) and individual decision rights, while Group 2 (*Utilitarians*) cares only about RBRs and utilitarian welfare criteria. In this framework, there is a clear potential for conflict between Group 2's two main fairness objectives: Rating anti-Black discrimination more harshly than similar acts of anti-White discrimination may not feel race-blind.[40] In this Section, we use the *race* treatment order effects documented in Section 2.4 to estimate the relative weight that members of Group 2 place on these two fairness criteria. Our identifying assumption is that respondents are not aware of their desire to make race-blind fairness assessments until they encounter a discriminatee from a second racial group, i.e. until they encounter a *switch* in the race treatment.

[39] In open-text answers, respondents who object to discrimination frequently say that it is wrong to base hiring decisions on race, statistical information, or tastes. These respondents frequently use words like "racism", "bigoted", "prejudice", "bias" and "stereotype" in their explanations. References to employer decision rights are essentially absent. See Appendix 6 for a detailed analysis of subjects' open-text responses.

[40] Since they are tolerant of both anti-White and anti-Black discrimination, Business Rights Advocates do not experience a similar conflict when the *race* treatment switches. To see this, consider for example a Business Rights advocate who said it was fair for a White business owner to accommodate his customers' anti-Black discriminatory tastes. In this case, race-blindness would be consistent with saying the same action is fair if the races were reversed, which is exactly how Business Rights Advocates behave in our survey.

To accomplish this goal, we proceed in three steps. First, we document that the race treatment order effects described in Section 2.4 are present in Group 2 but absent in Group 1. Second, we use a simple model of reporting behavior, combined with random assignment of the *race* treatment to interpret Group 2's order effects as a compromise between utilitarianism versus RBRs, and to estimate the relative weight Group 2 places on those two criteria when they conflict. Finally, for readers who may be skeptical of our *ex-post* Group 1 - Group 2 dichotomy, we replicate the preceding steps for the categories used in Section 4 of the paper: conservatives versus [moderates + liberals]. The results are very similar.

5.1 *Race* Treatment Order Effects are Absent in Group 1

In Section 2.4 we demonstrated the existence of race treatment order effects in our entire sample of respondents: Their Stage 2 fairness assessments depend on the *race* treatment they encountered in Stage 1. In Appendix 8.1, we show that there are no such order effects for Group 1: Regardless of the discriminatee race they encountered in Stage 1, members of Group 1 view discrimination as a little more fair than unfair (about +0.5 on a scale from -3 to +3) in Stage 2. Respondents from Group 2, on the other hand, exhibit a more pronounced version of the order effects we saw in the sample as a whole. Specifically, Group 2's Stage 2 fairness assessments of anti-Black discrimination are much milder if they encountered a White discriminatee (as compared to a Black discriminatee) in Stage 1 ($p = .009$).[41] Motivated by this fact, we restrict our attention to Group 2 respondents (*Utilitarians*) in the following sub-section, where we interpret Group 2's *race* treatment order effects as a compromise between the values they place on both utilitarianism and race-blindness.

5.2 A Trade-off between Utilitarianism and Race-Blind Rules?

As noted above, subjects who value both utilitarianism and race blindness (e.g. members of Group 2) face a conflict when they experience a switch in the *race* treatment. For example, in Stage 2, a White-to-Black treatment switcher needs to choose between assigning the same fairness rating they assigned to a White discriminatee in Stage 1 (race blindness), versus respecting their utilitarian desire to

[41] The race treatment a Group 2 subject received in Stage 1 does not have a statistically significant effect on the subject's ratings of anti-White discrimination in Stage 2.

object more strenuously to anti-Black than anti-White discrimination. Subjects who do not experience *race* treatment changes do not face this conflict.

To model this idea, we make the following assumptions:

Assumption 1:

> Subjects' Stage 1 assessments, B_i^1 and W_i^1 represent each respondent *i*'s "pure" utilitarian fairness ratings, B_i^* and W_i^*.

Assumption 1 seems reasonable because in Stage 1, respondents have not been asked to make any previous fairness assessments with which they might want to be consistent.

Assumption 2:

> In Stage 2, *race* treatment switchers care about two potentially conflicting things: reporting their pure utilitarian rating (B_i^* or W_i^*) for the *new* racial group, or making the same report they assigned to the other racial group in Stage 1 (being race-blind).[42]

Using this notation, in Stage 2 White-to-Black treatment switchers have the option of reporting their pure utilitarian rating of the group they now face in Stage 2 (thereby setting $B_i^2 = B_i^*$), assigning the same rating they assigned (to the other race) in Stage 1 (thereby setting $B_i^2 = W_i^1$), or reporting a weighted average of these two choices:

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

Where α is the weight placed on their utilitarian preference and $(1-\alpha)$ is the weight on their desire to make race-blind assessments. Our goal is to estimate α, but this is complicated by the fact that (unlike W_i^1 and B_i^2), B_i^* is not observed for White-to-Black treatment switchers.

To address this unobservability problem we take advantage of the fact our *race* treatments are randomly assigned. Thus, while B_i^* is not observed for W-to-B switchers (and W_i^* is not observed for B-

[42] Subjects' exposure to the Taste and Statistical treatments can change between Stages 1 and 2, but we abstract from that here since those treatments are randomly assigned and never appear to affect fairness assessments.

to-W switchers), their sample means $\bar{B}^*$ and $\bar{W}^*$ in any fixed population (such as Group 2) *are* observed for both groups of switchers from the mean Stage 1 responses of the subjects in their population who were randomly assigned to the other *race* treatment. We can therefore write:

$$\bar{B}^2 = \alpha\bar{B}^* + (1-\alpha)\bar{W}^* \quad (3)$$

where $\bar{B}^*$ and $\bar{W}^*$are sample means calculated from Stage 1 responses.

Similarly, for B-to-W switchers,

$$\bar{W}^2 = \alpha\bar{W}^* + (1-\alpha)\bar{B}^* \quad (4)$$

After restricting our sample to Group 2 respondents, Equations (3) and (4) can then be (separately) solved for α, yielding $\alpha = 0.49$ for the White-to-Black switchers and $\alpha = 0.68$ for the Black-to-White switchers.[43] Thus, W-to-B switchers' Stage 2 ratings of anti-Black discrimination place almost equal weight on race-blindness and utilitarianism. B-to-W switchers, on the other hand, act as if they place slightly more weight on utilitarianism than on race-blindness. The 95% percent confidence intervals for α are [0.243,0.800] and [0.405, 1.075] for W-to-B and B-to-W switchers, respectively. Thus, we cannot reject equal weight on both objectives ($\alpha = 0.5$) for either type of switcher. For W-to-B switchers, we can reject both α=0 and α=1, indicating strictly positive weight on both objectives. For B-to-W switchers, we reject α=0 but not α=1.

Summing up, the *race* treatment order effects we observe among our Group 2 (*Utilitarian*) respondents can be explained by a simple model that assumes these respondents value both the race-blind application of rules (RBRs) and utilitarian objectives. When these criteria conflict, i.e. when the respondent experiences a switch in the *race* treatment, respondents 'split the difference' about equally between these two objectives when making their fairness assessments.

[43] See Appendix 8.1 for the details of the calculations reported in this Section.

5.3 Replicating the Analysis by Political Orientation

In Appendix 8.2, we replicate the preceding analysis, splitting the sample by political orientation (conservatives versus moderates/liberals) with very similar results. Specifically, we show that race treatment order effects are absent among conservatives. Among moderates and liberals, they are stronger than in the full sample. Next, restricting attention to moderates and liberals, we use the same method to calculate α, the relative weight this group assigns to utilitarianism versus race blindness, The point estimates are $\alpha = 0.44$ for the White-to-Black switchers, and $\alpha = 0.62$ for the Black-to-White switchers, with confidence intervals [0.155,0.791] and [0.348, 1.033]. [44] As for Group 2, we conclude that moderates and liberals assign roughly equal weight to utilitarianism versus race-blindness when forced to choose between these two fairness criteria.

6. Robustness

One potential concern about the external validity of our results is the fact that our data were collected in September and October 2020, following a summer of civil unrest related to the murder of George Floyd on May 25, 2020. Together, these events led to a mainstream conversation on systemic racism in the U.S.; it seems reasonable to ask whether these events may have primed our respondents to answer our questions in unusual ways. To check for this possibility, Figure A2.1 presents online search trends for related keywords, including *Black Lives Matter* and *racism* during the spring and summer of 2020. These trends show that searches for these terms had diminished dramatically by the time of our survey, suggesting that this type of priming may not have been a significant issue for our respondents.

One striking result of our analysis is the large magnitude, statistical significance, and stability of the *justifiability* treatments, documented in Section 4.3: Respondents of all political orientations penalized the less- justifiable forms of discrimination (relative to more-justifiable forms) by the same amount, irrespective of the discriminatee's race. A possible concern with this result is the fact that the

[44] See Appendix 8.2 for the details underlying these calculations.

subjects always encounter both the *less* and *more* justifiable forms within each Stage (one right after the other), and that we draw subjects' attention to the sentences in the two scenarios that differ from each other. Thus, subjects may have taken special care to how they rank the fairness of these two types of scenarios, with respect to each other. To address this issue, Appendix A5.1 replicates Tables 1 and 2 using only the first individual scenario each respondent encountered. The results are almost identical to our main estimates, suggesting that subjects' desires to maintain a consistent ranking of the two types of scenarios are not responsible for this finding.

Figure 8 illustrated some strong and unexpected relationships among subjects' beliefs about relative opportunities (BROs), subjects' political orientation, the race of the discriminatee, and subjects' fairness assessments. Together, these relationships were starkly inconsistent with the belief-based utilitarian model of fairness. To probe the robustness of these results to the fact that Figure 8 combines moderates and liberals into a single group, Figure A5.3 replicates Figure 8, showing separate results for moderates versus liberals. Consistent with other results in the paper, these two groups exhibit very similar response patterns, both of them differing substantially from conservatives' patterns.

In Section 5 of the paper, we used *race* treatment order effects to estimate the relative weight a subset of our subjects assign to utilitarian and race-blind fairness criteria. These order effects could, however, be caused by a particular form of experimenter demand effects that seems quite plausible in our context. To see this, consider the following possibility: If respondents encounter the Black treatment in Stage 1, they assume that we (the experimenters) are either moderate or liberal. Then – to please us – the respondents provide Stage 1 fairness assessments that are typical for moderates and liberals (i.e. discrimination against Black applicants is unfair, and more unfair than discrimination against White applicants). On the other hand, if respondents encounter the White treatment in Stage 1, they assume the experimenters are conservative and provide Stage 1 answers that are typical for conservatives (i.e., discrimination against both Black and White applicants is neutral or fair). Finally, respondents who encounter a change in the *race* treatment between Stages 1 and 2 update their priors to become uncertain about the experimenters' politics and moderate their fairness assessments accordingly. Together, these patterns could account for exactly the type of race treatment order effects we observe, where (for

example) respondents' Stage 2 opposition to anti-Black discrimination is reduced if they encountered anti-White discrimination in Stage 1.

In Appendix 7, we provide highly suggestive evidence against this possibility based on the idea that subjects who want to please the experimenters should tailor not just their fairness assessments but also their answers to other survey questions to achieve the same end. Of particular interest in this regard are the subjects' assessments of Black peoples' relative economic opportunities (BRO), and potentially even subjects' reported political orientations (all elicited in Stage 3 of the survey). For example, suppose a subject encountered the White treatment in both Stage 1 and 2 of the survey. Under our assumptions about experimenter demand effects, this should send a strong signal that the experimenters are conservatives. To please us, we would then expect the subjects to report that Black people have a higher level of relative economic opportunity, and perhaps even to shade their own reported political leanings in a more conservative direction on our seven-point scale.

In Appendix 7, we examine whether subjects' responses to these Stage 3 questions depend on the *race* treatments they received in Stages 1 and 2, and find no such effects: Specifically, subjects' BRO assessments, stated political party preferences, and reported left-right leaning are highly stable with respect to the *race* treatments they encountered earlier in the experiment. We conclude that experimenter demand effects of this type are probably not responsible for the *race* treatment order effects we observe.[45] While not conclusive, the stability of subjects' Stage 3 responses to previously-encountered race treatments also suggests the experimenter demand effects are likely not responsible for the strong and robust *justifiability* effects we estimate.

A broader concern that could affect the validity of all our results is the fact that our sample of MTurk respondents was not representative of adult Americans on a number of key dimensions, including

[45] Additional evidence against this demand-effects hypothesis is the fact, documented in Appendix 8.2, that race treatment order effects are absent among conservatives. For experimenter demand effects to explain our results in Section 5, these demand effects must *only* be present among moderates and liberals. In other words, moderates and liberals should want to please an experimenter they perceive as moderate or liberal, but conservatives must have no such desire to please a conservative experimenter. In contrast, the fairness reporting model in Section 5 has a 'built in' explanation for conservatives' lack of order effects: Conservatives do not value utilitarian objectives, so they experience no conflict between utilitarianism and the fairness criteria they care about.

age and education (see Appendix 2 for details). While our small sample size limits what we can do to address this issue, Appendices 9 and 10 replicate all our main results (Figures 2-8) two different ways. First, Appendix 9 uses the 2019 American Community Survey to re-weight our MTurk responses by the relative prevalence of our respondents in 24 cells, defined by gender, race, education, and age. All the main patterns discussed in the paper are replicated, with one small exception: the weak positive association between BRO and the fairness of anti-Black 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. Second, Appendix 10 replicates Figures 2-8 using weights derived from the 2020 General Social Survey (GSS) which are based only on a 7-point political leaning scale (i.e., extremely conservative, conservative, slightly conservative, moderate, slightly liberal, liberal, and extremely liberal) that is asked in a very similar way to our survey.[46] Despite significant differences in the political mix of the two surveys, all the main results are replicated.[47]

As documented in Appendix P, one of the main differences between the results reported in our paper and the methods proposed in our pre-analysis plan (PAP) involves how our main outcome variable (fairness) is measured. While the PAP proposed a standardized measure of fairness, we realized that a standardized fairness measure centered at the mean fairness level across all respondents and treatments would obscure meaningful cardinal information, relative to the measure we decided to use. Our measure is centered at a conceptually meaningful level --"neither fair nor unfair"-- and our measure's integer values correspond directly to the seven fairness categories respondents could choose from. We made a similar choice with respect to our measure of Black people's relative opportunities (BRO), centering it at "roughly equal opportunities" rather than the sample mean. To see if these decisions had any effects on

[46] In our GSS re-weighting exercise we ignored the difference in the wording of the middle political leaning category between the GSS and our survey. As noted, this might exaggerate the difference between the two surveys, so it should give us an upper bound on the effects of re-weighting. 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 extreme and imprecise weights. The ACS does not ask questions about political orientation or party preference.

[47] The one exception noted with the ACS weights in Appendix 9 does not occur here.

the main relationships established in the paper, Appendix 11 replicates Figures 2-8 using standardized fairness and BRO measures. There is no meaningful difference.

A common critique of non-incentivized survey experiments like ours is that subjects have little incentive to answer the questions thoughtfully, leading to results that are noisy, or simply different from what the same person might offer if they took more time to think about the question. We took a number of precautions to prevent this (see Section 2.3), but it remains possible that many of our respondents gave careless answers that differ from what a thoughtful person would choose. To assess this possibility, Appendix 12 replicates all our main results (Figures 2-8) for 'thoughtful' respondents only, where 'thoughtful' is defined as taking more than the median amount of time to complete the survey. No meaningful differences were evident.

A final important concern --which affects virtually all tests of statistical hypotheses-- is the extent to which the hypotheses were selected after a preliminary analysis of the data. To address this concern, we posted a registered pre-analysis plan (PAP) before launching our survey. The relationships between the analyses proposed in the PAP and the hypotheses tested in our survey are described in detail in Appendix P. Briefly, Appendices P1-P3 together comprise a "populated PAP" which reports the results of the exact tests specified in the PAP. Appendix P4 summarizes the relationship between the PAP and the paper. In a little more detail, Appendix P4 shows that the following key analyses in the paper were declared in advance: all the descriptive "facts" presented in Section 3; all four theoretical models of discrimination described in Sections 4.1-4.4 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 question order effects to learn about respondents' preferences for race-blindness (see Appendix P2.5). Appendix P4 also describes the five most important ways in which our main analyses in the paper differ from the PAP. These are all relatively minor, and the populated PAP results in Appendices P1-P3 strongly suggest they do not matter. Finally, Appendix P4 notes that there are only two PAP hypothesis tests that we decided *not* to include in the main paper and discusses our motivations for those decisions.

7. Discussion

Inspired by a rapidly growing literature on the perceived fairness of pay and income inequality, and by a large literature on discrimination, we have used an MTurk survey to elicit Americans' assessments of the fairness of canonical examples of statistical and tasted-based racial discrimination. We find, first of all, that conservative respondents are more accepting of discriminatory actions than moderate and liberals. Second, while distinguishing between statistical and taste-based discrimination has been of considerable interest to economists, whether discrimination is motivated by (someone's) tastes or by statistical reasons is not a reliable predictor of assessed fairness. Third and in contrast, respondents of all political leanings *do* care about other aspects of the motivation behind a discriminatory act. Specifically, our respondents agree that acting on one's *own* tastes is less fair than accommodating others' tastes, and that using imprecise or inaccurate statistical information is less fair than precise information. Indeed, respondents of all political leanings penalize these less-justifiable actions by the same amount, and do so regardless of discriminatee race, suggesting a broad area of common ground in how Americans react to different discriminatory actions. Fourth, another important partisan difference is that only moderates and liberals consider the race of the discriminatee when assessing the fairness of a discriminatory act.

Comparing the preceding findings with four pre-registered models of how respondents might make fairness assessments, we find that two of those models – in-group bias models and belief-based utilitarianism –conflict with several key patterns in our data. Using open-text data to identify an unanticipated rationale underlying some subjects' fairness assessments, we propose an *ex post* interpretive framework with two equally-sized political groups and three models of fairness – simple utilitarianism, race-blind rules (RBRs), and employer decision rights – that can account for most of the fairness patterns we observe. In this model, both political groups value using a set of race-blind rules to compare the fairness of different types of discriminatory actions. One group, who we call "Business Rights Advocates" and are mostly conservative, also value employer decision rights. The other group, "Utilitarians" is a large subset of moderates and liberals. They value utilitarian fairness criteria in addition to RBRs, but not employer decision rights. When their utilitarian and RBR objectives conflict – as when

they experience a change in the experiment's *race* treatment – we estimate that members of Group 2 place about equal weight on these two objectives.

While our main objective in this paper has been to understand when and why people view discriminatory actions as fair or unfair, our findings may also have some implications for both managerial and public policy. In a management or human resources context, our findings suggest that workers' perceptions of the fairness of policies or actions with disparate impacts on racial groups are likely to depend on the precise motivations or circumstances surrounding those policies or actions.[48] Interestingly, since our data show that 'reasons matter' to members of all political groups, our evidence suggests that employers may reap wide benefits from transparent, rules-based recruitment and pay policies that provide clear justifications for any decisions that have disparate racial impacts.

In terms of public policy, our study suggests the potential for substantial political headwinds for certain anti-discrimination policies. While acts of anti-Black discrimination are viewed as unfair by a majority (63.1%) of our sample, the rest of our respondents view the discriminatory actions depicted in our scenarios as either neutral or fair, regardless of the race of the discriminatee. Our results suggest that this group of respondents is likely to resist policies that interfere with employers' decision rights, even when those hiring decisions represent canonical examples of taste-based and statistical discrimination on the basis of race. That said, our analysis also suggests two types of situations in which conservative Americans might be more receptive to policies that equalize racial opportunities. One such situation is where a clear rule has *not* been applied in a race-blind way; in these cases, *restoring* race-blindness should have broad appeal given our results. Second, we show that respondents of all political leanings react more negatively to race-based actions that were taken for less-justifiable reasons, like personal animus and low-quality evidence. Antidiscrimination policies that target these types of behaviors may thus be better received than other policies.

[48] In this sense our findings complement existing evidence that the motivations behind underlying pay differentials (Frank, 1984; Charness and Kuhn, 2007; Gartenberg and Wulf, 2017; Mas, 2017; Breza et al. 2017) and layoffs (Charness and Levine, 2000) have a large effect on their acceptability to workers.

Our results in this paper are subject to some important *caveats* and leave some important questions unanswered. One important *caveat* is that all our results are *limited to the range of actions our scenarios depict*. Thus, for example, it seems likely that more *consequential* discriminatory actions (like being fired from a job or convicted of a crime), and less *justifiable* actions (such as ones based on racial hatred) would probably elicit stronger negative responses from respondents than we see. We might also see stronger, negative reactions, for example, to hiring scenarios in which the discriminatee is *more* qualified than his co-applicant. (We restrict attention to equally qualified applicants). Another limitation is that our scenarios are confined to a particular type of firm: sole proprietorships. We chose this context because it ensures that the recruiter has total control of the hiring decision and experiences its full financial consequences.[49] In larger firms, recruiters might not bear the full costs of indulging their own tastes or using lower-quality information. The strong and widespread support we see for *employer rights* among our respondents might also be more muted when the recruiter is an employee of a large firm.

While we have more than enough statistical power to test our pre-registered hypotheses, we also acknowledge that we lack statistical power to answer two important questions. First, does the discriminatee race effect really reverse sign (relative to the sample as a whole) among White conservatives? If White conservatives truly object more strongly to anti-White and anti-Black discrimination, this would complicate our description of conservatives, in general, as valuing race-blindness. Second, given our small sample, our data cannot shed much light on which factors explain non-White respondents' fairness assessments. Finally, we remind readers that the fairness assessments we elicit are not necessarily the same as the *actions* our respondents might take in real-world situations similar to our scenarios. For example, a business owner might choose to accommodate the discriminatory tastes of her customers while still experiencing that action as unfair.[50] That said, given the extensive evidence that people value fairness (e.g. Card et al. 2012; Cullen and Perez-Truglia 2018; Dube et al,

[49] In this respect, we follow Becker's (1971) classic exposition of employer taste-based discrimination: Becker's 'employers' make all of a firm's decisions (including hiring) and receive all the profits generated from the firm's operations. Assigning fairness ratings to our scenarios would be both more complex and more interesting if, for example, recruiters are balancing their personal assessments of what is best against company policies.

[50] That said, we note that on average our respondents rated this scenario as slightly more fair than unfair, suggesting that our respondents' real-world actions might indeed coincide with their fairness assessments in this case.

2019) our paper quantifies, for the first time, the fairness *costs* our subjects associate with taking different types of discriminatory actions.

Given all the above limitations, we view our results in this paper as a first step in understanding when and why ordinary people view discriminatory actions as unfair. One of many questions that could fruitfully be addressed in extensions of our work is the effects of the discriminatee's *individual* income (and opportunities) on respondents' fairness assessments. (Respondents' reactions to rich discriminatees from low-income groups could shed further light on utilitarianism, for example.) Other applications include different contexts, such as housing markets, credit markets, and judicial decisions; different discriminatee groups (such as gender, age, sexual orientation, political orientation, age, criminal and credit history); different social and psychological contexts *in the scenario* (for example, is the hypothetical action seen by hypothetical observers?; is the act depicted as conscious versus unintended?); different decision environments *for the respondent* (such as priming, cognitive depletion, audience effects, and personal exposure to previous discriminatory actions); and discrimination that is *embedded in laws and institutions* (as opposed to an individual's actions).

References

Abeler, Johannes, Sebastian Kube, Steffen Altmann, and Matthias Wibral. 2010. "Gift Exchange and Workers' Fairness Concerns: When Equality is Unfair." *Journal of the European Economic Association* 8 (6): 1299-1324.

Alesina, Alberto, and Eliana La Ferrara. 2005. "Preferences for Redistribution in the Land of Opportunities." *Journal of Public Economics* 89: 897-931.

Alesina, Alberto, Armando Miano and Stefanie Stantcheva. 2020 "The Polarization of Reality" *American Economic Review Papers and Proceedings* 110: 324-328

Alesina, Alberto, Matteo F. Ferroni, and Stefanie Stantcheva. 2021. "Perceptions Of Racial Gaps, Their Causes, And Ways To Reduce Them" NBER working paper no. 29245

Almås, Ingvild & Cappelen, Alexander & Tungodden, Bertil. (2020). "Cutthroat Capitalism versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking than Scandinavians?" *Journal of Political Economy*, Volume 128, Number 5.

Andreoni, James, Deniz Aydin, Blake Barton, B. Douglas Bernheim and Jeffrey Naecker. 2020. "When Fair Isn't Fair: Understanding Choice Reversals Involving Social Preferences." *Journal of Political Economy* 128(5): 1673-1711.

Arechar, Antonio A. Simon Gachter, and Lucas Molleman. 2018. "Conducting Interactive Experiments Online." *Experimental Economics* 21: 99-131.

Auspurg, Katrin, Thomas Hinz, and Karsten Sauer. 2017 "Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments" *American Sociological Review* Vol 82, Issue 1.

Barr, Abigail, Tom Lane and Daniele Nosenzo. 2018. "On the Social Inappropriateness of Discrimination." *Journal of Public Economics* 164: 153–164.

Becker, Gary S. 1971. *The Economics of Discrimination* (second edition) Chicago: University of Chicago Press.

Bertrand, Marianne, and Esther Duflo. 2017. Field Experiments on Discrimination. In Banerjee, Abhijit Vinayak and Esther Duflo, eds. *Handbook of Economic Field Experiments*, vol. 1. Chapter 8 (pages 309-393) Also available as: NBER Working Paper 22014, 2016.

Bohren, J. Aislin, Kareem Haggag, Alex Imas, and Devin G. Pope. 2019. "Inaccurate Statistical Discrimination." NBER Working Paper No. 25935.

Bracha, Anat, Uri Gneezy, and George Loewenstein. 2015. **"**Relative Pay and Labor Supply." *Journal of Labor Economics* 33 (2): 297-315.

Breza, Emily, Supreet Kaur, and Yogita Shamdasani. 2017. "The morale effects of pay inequality." *The Quarterly Journal of Economics 133(2)*: 611-663.

Bruhin, Adrian, Ernst Fehr, and Daniel Schunk. 2019. "The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences." *Journal of the European Economic Association* 17(4): 1025–1069.

Cain, Glen G. 1986. "The Economic Analysis of Labor Market Discrimination: A Survey." In *Handbook of Labor Economics*, Vol. 1, edited by O. Ashenfelter & R. Layard, 693-781. Elsevier.

Cappelen, Alexander W., Ranveig Falch and Bertil Tungodden 2019 "The Boy Crisis: Experimental Evidence on the Acceptance of Males Falling Behind" NHH Department of Economics Discussion Paper No. 06/2019.

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez. 2012. "Inequality at work: The effect of peer salaries on job satisfaction." *American Economic Review* 102(6): 2981-3003.

Charness, Gary., Till Gross, and Christopher Guo. 2015. "Merit Pay and Wage Compression with Productivity Differences and Uncertainty." *Journal of Economic Behavior & Organization* 117: 233-247.

Charness, Gary and David I. Levine. 2000. "When Are Layoffs Acceptable? Evidence from a Quasi-Experiment." *Industrial and Labor Relations Review* 53(3): 381-400.

Charness Gary and Peter Kuhn. 2007. "Does Pay Inequality Affect Worker Effort? Experimental Evidence". *Journal of Labor Economics* 25(4): 693-724.

Chen, Y. and Li, S. X. 2009. "Group identity and social preferences." *American Economic Review* 99(1): 431–457.

Cohn, Alain, Ernst Fehr and Lorenze Götte. 2014. "Fair Wages and Effort Provision: Combining Evidence from a Choice Experiment and a Field Experiment." *Management Science*, 61(8): 1777-1794.

Cullen, Zoe B. and Bobak Pakzad-Hurson. 2017. "Equilibrium Effects of Pay Transparency." unpublished paper, Harvard Business School.

Davidai, S. and J. Walker (2021). Americans Misperceive Racial Disparities in Economic Mobility. *Personality and Social Psychology Bulletin*, 01461672211024115.

Everett, Jim A. C., Nadira S. Faber, and Molly Crockett. (2015). Preferences and beliefs in ingroup favoritism *Frontiers in Behavioral Neuroscience,* volume 9.

Feess, E., Feld, J., and Noy, S. (2021). "People Judge Discrimination Against Women More Harshly Than Discrimination Against Men - Does Statistical Fairness Discrimination Explain Why?" *Frontiers in psychology*, *12*, 675776.

Fehr, Dietman, Hannes Rau, Stefan T. Trautmann and Yilong Xu. 2021. "Fairness Properties of Compensation Schemes". Unpublished paper, University of Heidelberg.

Fong, C. M. and E. F. Luttmer (2009). What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty. *American Economic Journal: Applied Economics* 1 (2), 64-87

Fong, C. M. and E. F. Luttmer (2011). "Do fairness and race matter in generosity? Evidence from a nationally representative charity experiment" *Journal of Public Economics* 95 (5), 372-394.

Frank, Robert H. 1984. "Are Workers Paid Their Marginal Products?" *American Economic Review* 74(4): 549–571.

Gartenberg, Claudine, and Julie Wulf. 2017."Pay Harmony? Social Comparison and Performance Compensation in Multibusiness Firms." *Organization Science* Vol. 28(1): 39-55.

Griggs v. Duke Power Co. 1971. 401 U.S. 424.

Haaland, I. and C. Roth (2021). Beliefs about racial discrimination and support for pro-black policies. *Review of Economics and Statistics*, forthcoming.

Jasso, Guillermina and Peter H. Rossi (1977) Distributive Justice and Earned Income *American Sociological Review* Vol. 42, No. 4 (Aug. 1977), pp. 639-65.

Jasso, Guillermina, Robert Shelly and Murry Webster 2019 "How impartial are the observers of justice theory?" *Social Science Research* 79: 226-246.

Kraus, M. W., I. N. Onyeador, N. M. Daumeyer, J. M. Rucker, and J. A. Richeson (2019). The Misperception of Racial Economic Inequality. *Perspectives on Psychological Science* 14 (6), 899–921.

Kraus, M. W., J. M. Rucker, and J. A. Richeson (2017). Americans misperceive racial economic equality. *Proceedings of the National Academy of Sciences* 114 (39), 10324–10331.

Kuhn, Peter and Trevor Osaki. 2020. "When is Discrimination Unfair?" AEA RCT Registry. September 22.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez and Stefanie Stantcheva. 2015 "How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review* 105(4): 1478-1508.

Krupka, Erin L. and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association* 11(3): 495– 524.

Lefgren, Lars J., David Sims and Olga Stoddard. 2016. "Effort, luck, and voting for redistribution". *Journal of Public Economics* 143: 89-97.

Hedegaard, Morten Størling and Jean-Robert Tyran. 2018. "The Price of Prejudice" *American Economic Journal: Applied Economics* 10(1): 40–63.

Lippens, Louis, Stijn Baert, and Eva Derous. 2021. "Loss Aversion in Taste-Based Employee Discrimination: Evidence from a Choice Experiment*." IZA discussion paper* no. 14438.

Mas, Alexandre. 2017. "Does Transparency Lead to Pay Compression?" *Journal of Political Economy* 125(5): 1683-1721.

Oprea, Ryan and Sevgi Yuksel 2021. "Social Exchange of Motivated Beliefs" *Journal of the European Economic Association*, forthcoming.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. (2021) Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54. pages 1643–1662.

Sauer, Carsten. 2020 "Gender Bias in Justice Evaluations of Earnings: Evidence From Three Survey Experiments" *Frontiers in Sociology* 7(5):22.

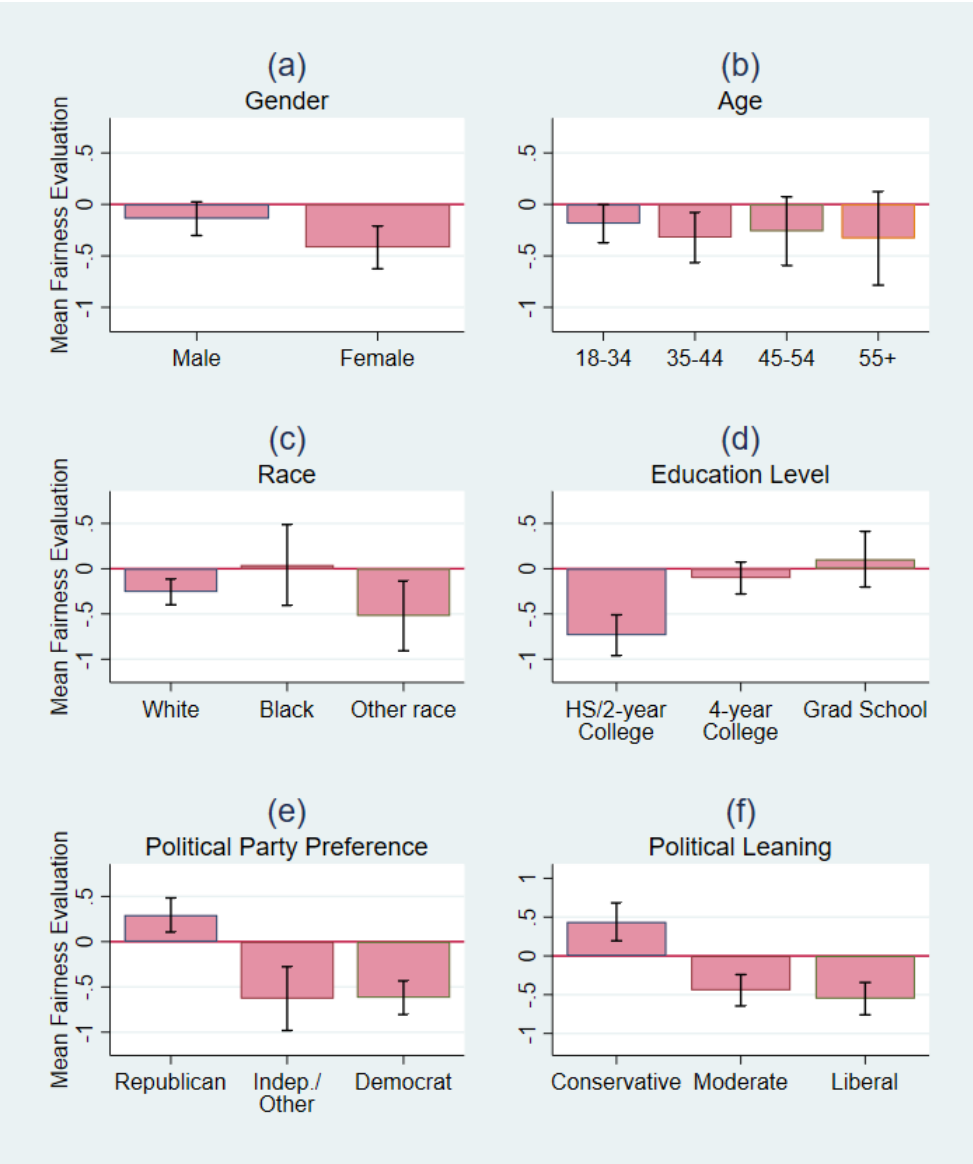
Schildberg-Hörisch, Hannah, Marco A. Schwarz, Chi Trieu, and Jana Willrodt 2022 "Perceived Fairness and Consequences of Affirmative Action Policies" CESifo working paper no.10198

Stantcheva, Stefanie. 2021. Understanding Tax Policy: How do People Reason? *Quarterly Journal of Economics* 136(4): 2309–2369.

Tilcsika, András 2021. "Statistical Discrimination and the Rationalization of Stereotypes" *American Sociological Review* 86(1): 93–122.

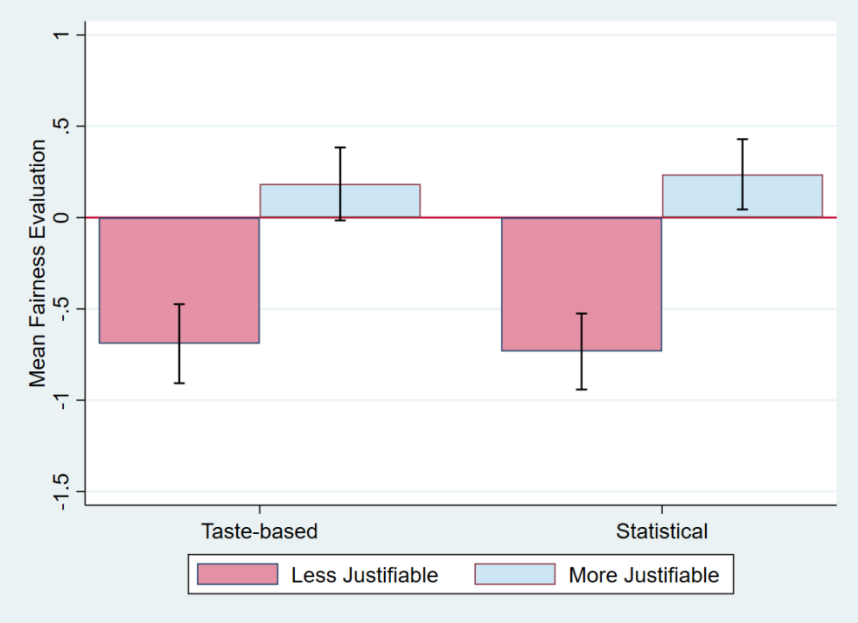
Figures

Figure 1: Mean Fairness of Discriminatory Actions by Respondent Characteristics



Notes: Fairness is measured on a scale from -3 ("very unfair") to 3 ("very fair"), where 0 was "neither fair nor unfair." This figure is based on only Stage 1 observations. 95% confidence intervals are shown. The *p*-values below are clustered by respondent.		
a) Gender: Males vs. Females = 0.037 **b) Age:** Ages 18-34 vs. 35-44 = 0.368 Ages 35-44 vs. 45-54 = 0.766 Ages 45-54 vs. 55+ = 0.805 Ages 18-34 vs. 55+ = 0.560	**c) Respondent Race:** White vs. Black = 0.204 Black vs. Other = 0.058 White vs. Other = 0.197 **d) Education level:** HS/2-year College. vs. 4-year College = 0.000 4-year College vs. Grad School = 0.246 Grad school vs. HS/2-year College = 0.000	**e} Political party preference** Republicans vs. Independents = 0.000 Independents vs. Democrats = 0.961 Democrats vs. Republicans – 0.000 **f) Political leaning:** Conservatives vs. Moderates = 0.000 Moderates vs. Liberals = 0.463 Liberals vs. Conservatives = 0.000

Figure 2: Fairness Ratings by Type of Discrimination and *Justifiability*



p-values:

Less- versus more justifiable treatments:

Overall: p=.000

Within taste-based: p=.000

Within statistical: p=.000

Taste versus Statistical Discrimination:

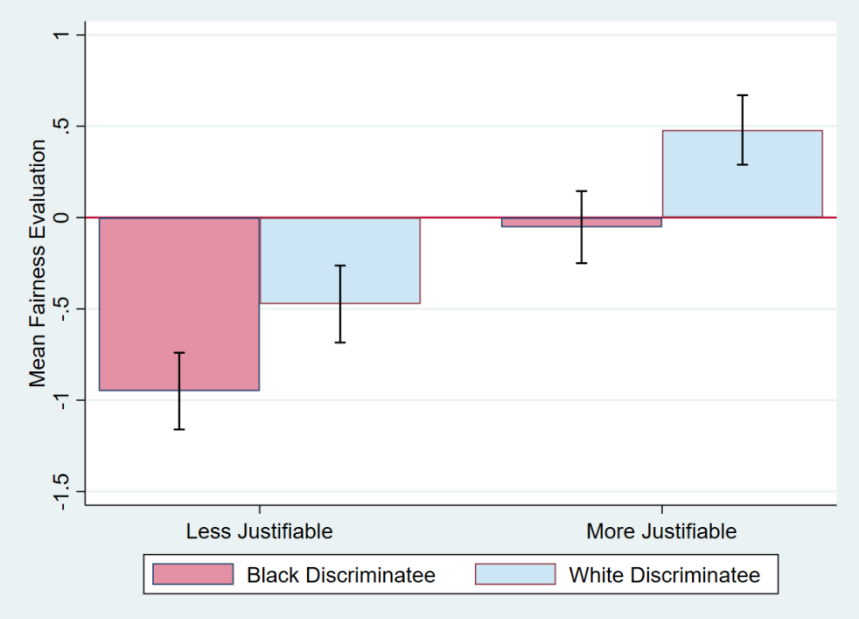
Overall: p=.971

Within Less-Justifiable: p=.779

Within More-Justifiable: p=.710

Note: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. *p*-values are clustered by respondent.

Figure 3: Fairness by *Justifiability* and Discriminatee Race



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

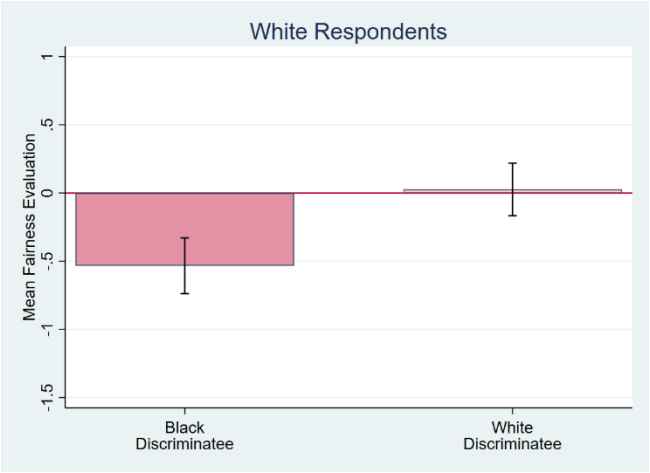
Within White Discriminatees: p=.000

Note: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. *p*-values are clustered by respondent.

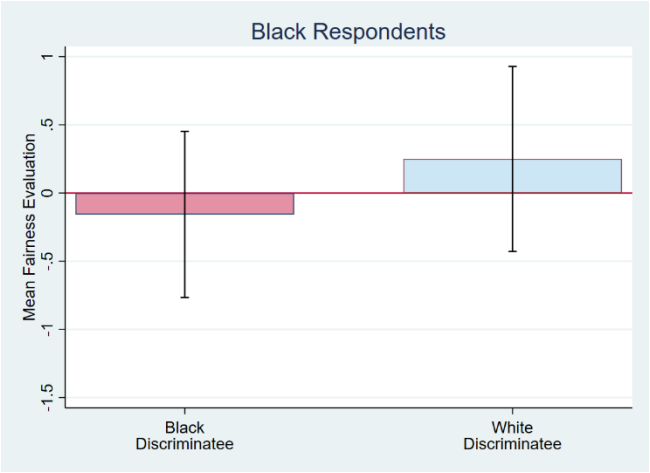
Within Black Discriminatees, less-justifiable scenarios are 0.898 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.953 units less fair. A test for equality of the Less versus More *Justifiability* Gap between the Black and White treatment yields p = .679.

Figure 4: Fairness Ratings by Respondent Race and Discriminatee Race

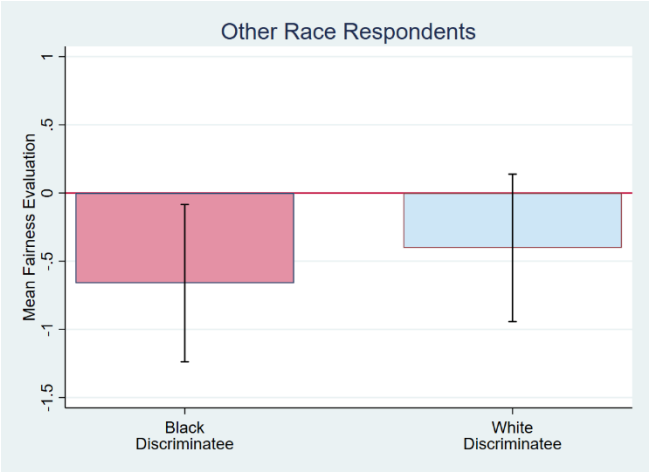
(a)



(b)



(c)



p-values:

Black versus White Treatment:

Overall: p=.000

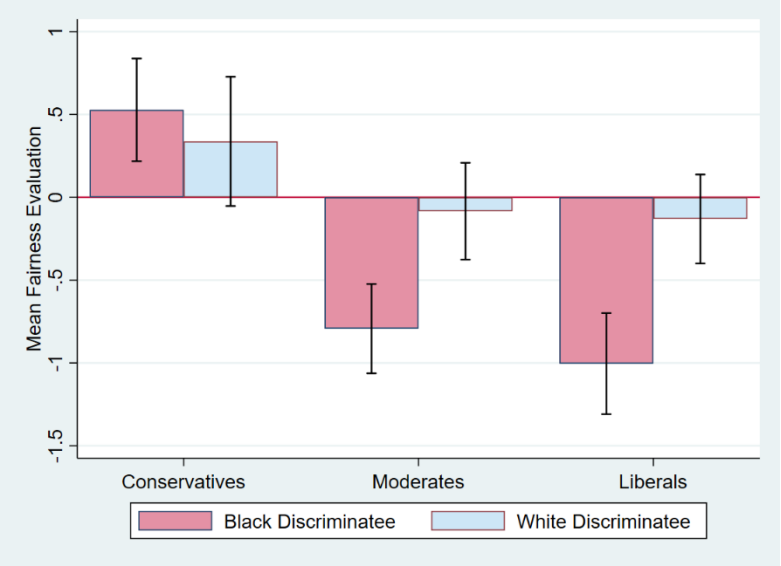
Within White respondents: p=.000

Within Black respondents: p=.365

Within Other respondents: p=.506

Note: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. 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 5: Fairness Ratings by Political Orientation and Discriminatee Race



p-values:

Black versus White Treatment:

Overall: *p*=.000

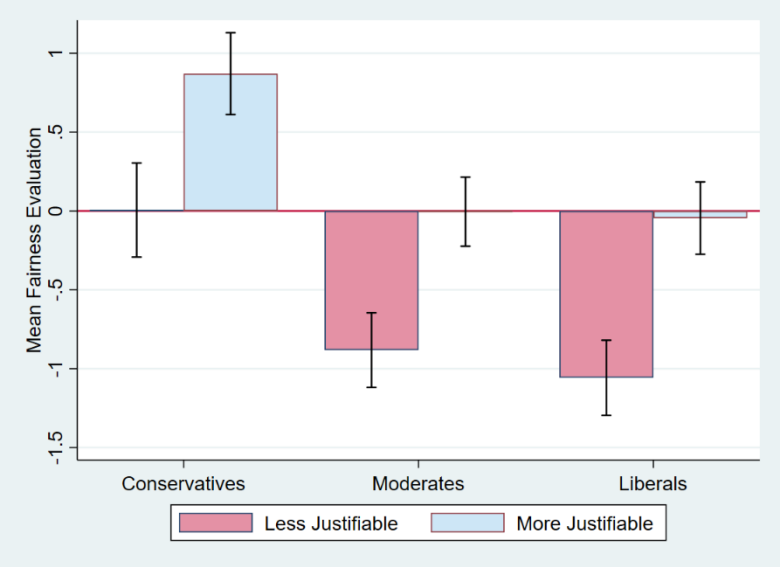
Within Conservatives: *p*=.448

Within Moderates: *p*=.000

Within Liberals: *p*=.000

Note: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. 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 6: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent's Political Leaning



p-values:

Less versus More Justifiable Treatment:

Overall: *p*=.000

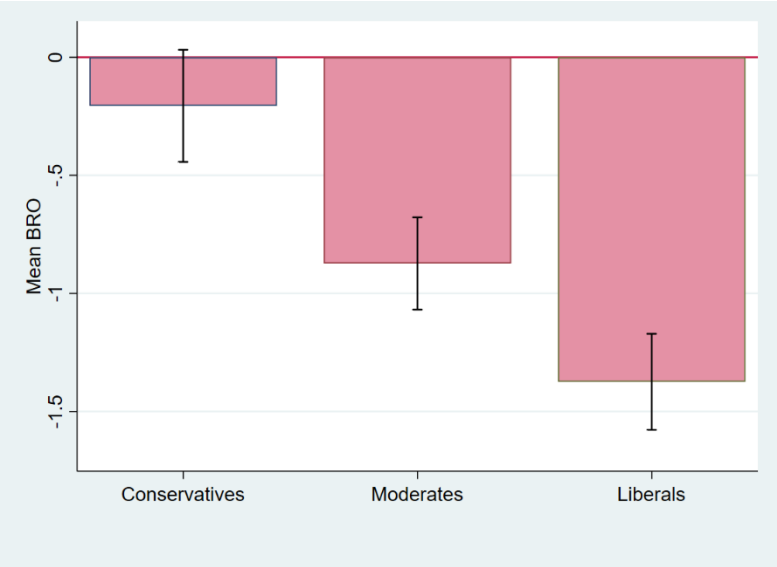
Within Conservatives: *p*=.000

Within Moderates: *p*=.000

Within Liberals: *p*=.000

Note: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. 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 7: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning



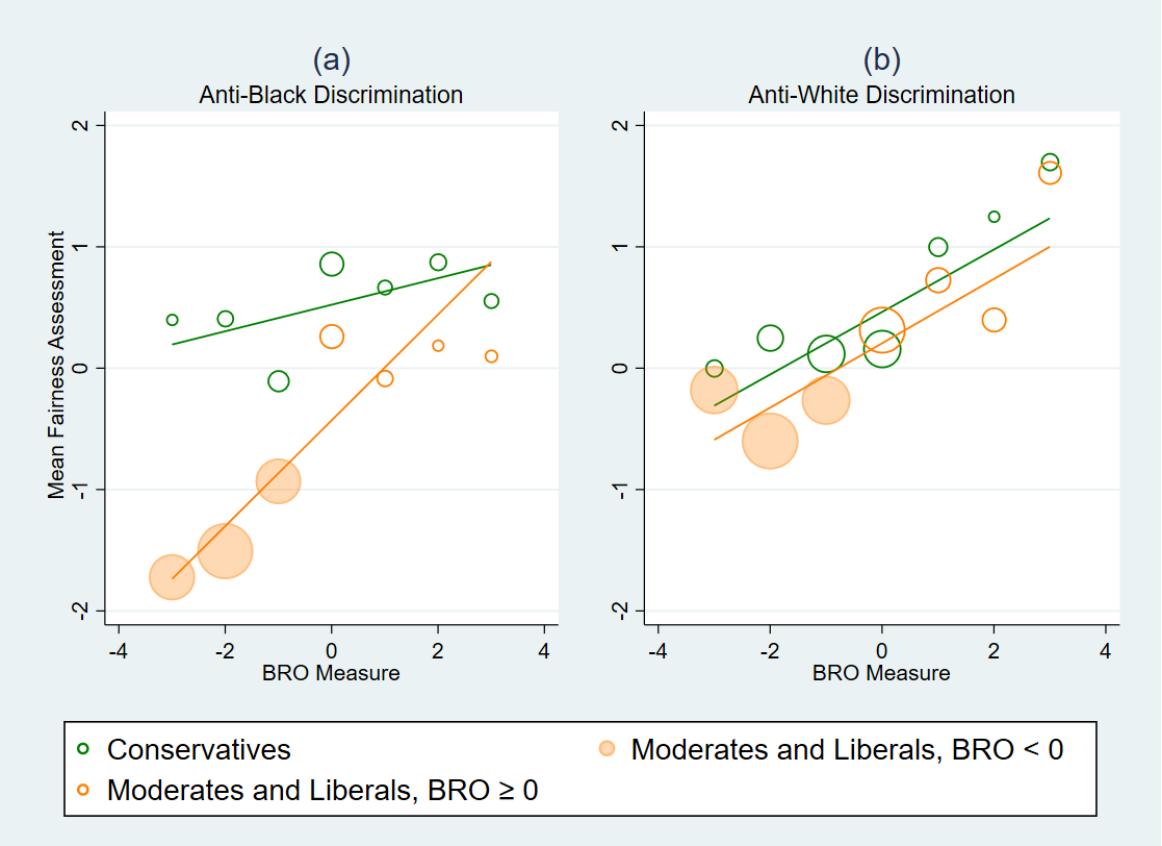
p-values:

Conservatives vs. Moderates: p=.000
Moderates vs. Liberals: p=.001
Liberals vs. Conservatives p=.000

Note:

BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more). Figure is based on Stage 1 observations only. 95% confidence intervals are shown. All *p*-values are clustered by respondent. A test for equality of BRO across all three political groups yields p = .577.

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



Note: 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.

- Panel (a), Discrimination against Black Applicants
  - For Conservatives: slope = 0.109, $p = .218$
  - For Moderates and Liberals, slope = 0.436, $p = .000$
- Panel (b), Discrimination against White Applicants
  - For Conservatives: slope = 0.257, $p = .094$
  - For Moderates and Liberals, slope = 0.265, $p = .000$

Political leaning subsamples for anti-Black discrimination:

Conservatives vs. Mods-Libs, BRO = -3 only ($p = .000$)

Conservatives vs. Mods-Libs, BRO = -2 only ($p = .000$)

Conservatives vs. Mods-Libs, BRO = -1 only ($p = .658$)

Conservatives vs. Mods-Libs, BRO = 0 to +3 combined, only ($p = .000$)

Appendix (for online publication)

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."

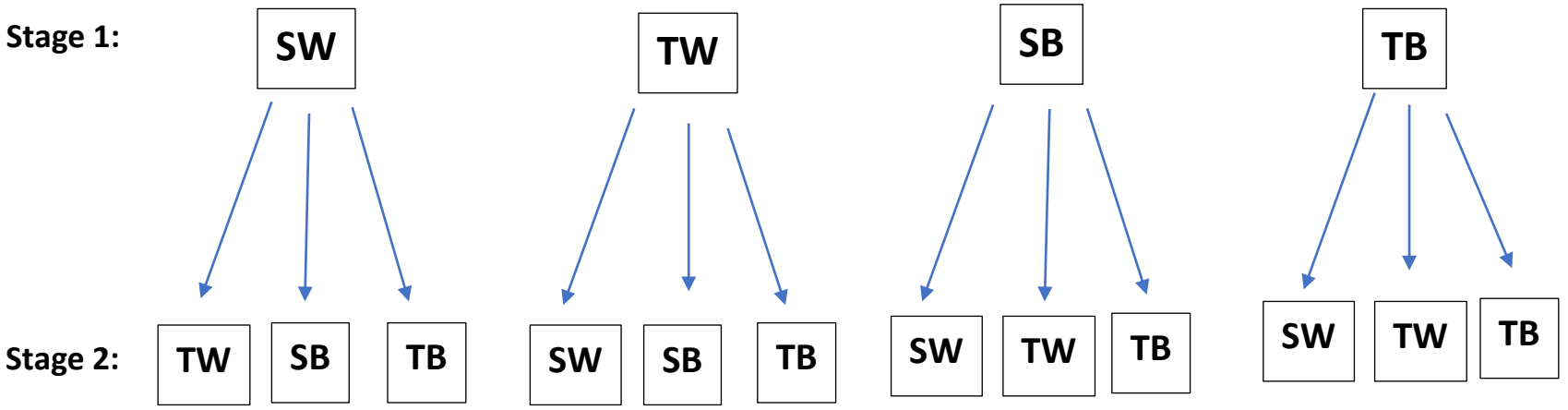
Please do not exit the survey from this page. You must click on the "next button" to reach the "end of survey" page so that your responses are recorded. This button will appear in a few seconds.

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.[51] 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.

[51] 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

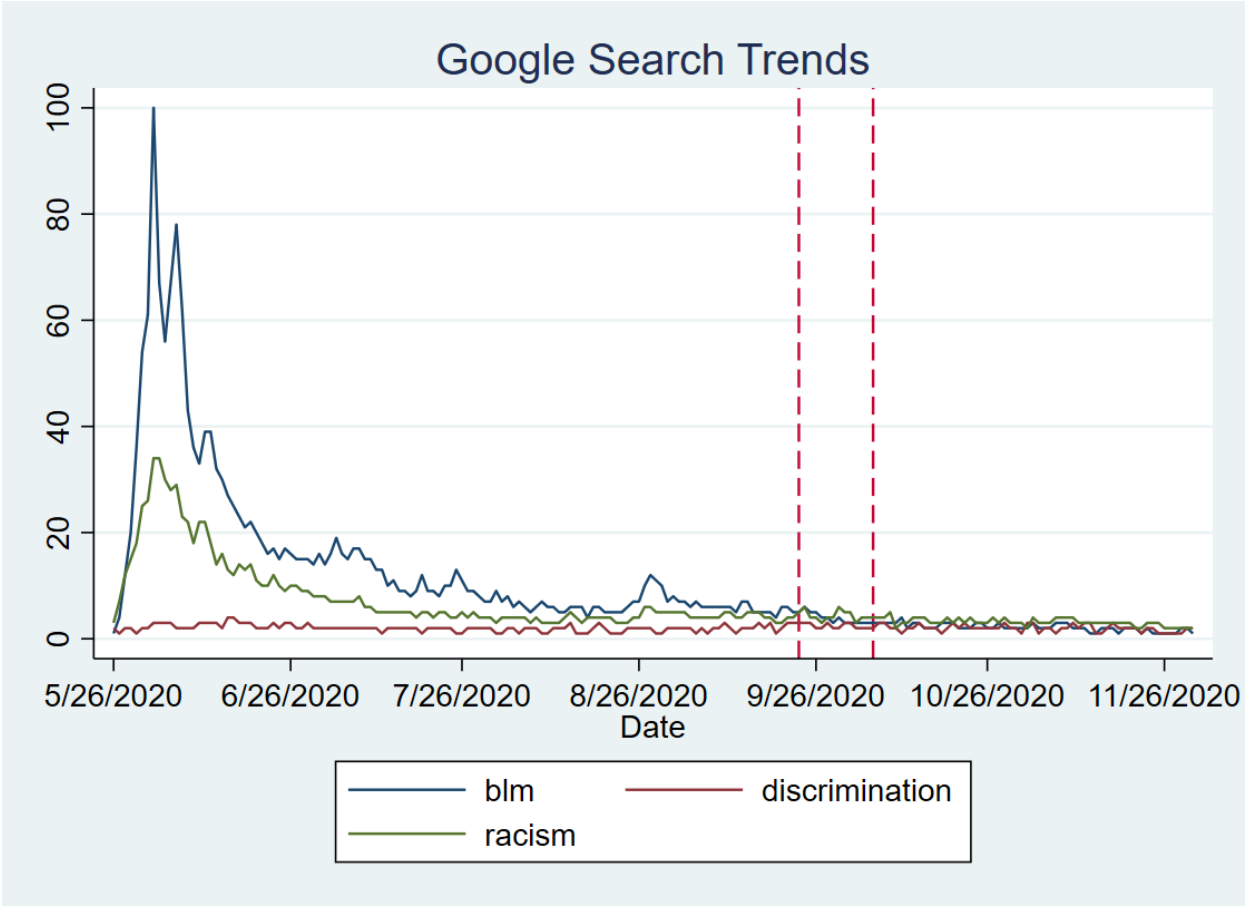
Notes: Column 1 contains the percentage of respondents across U.S. census regions and their respective divisions. Column 2 contains these percentages for the 2019 American Community Survey (ACS) sample for comparison.

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



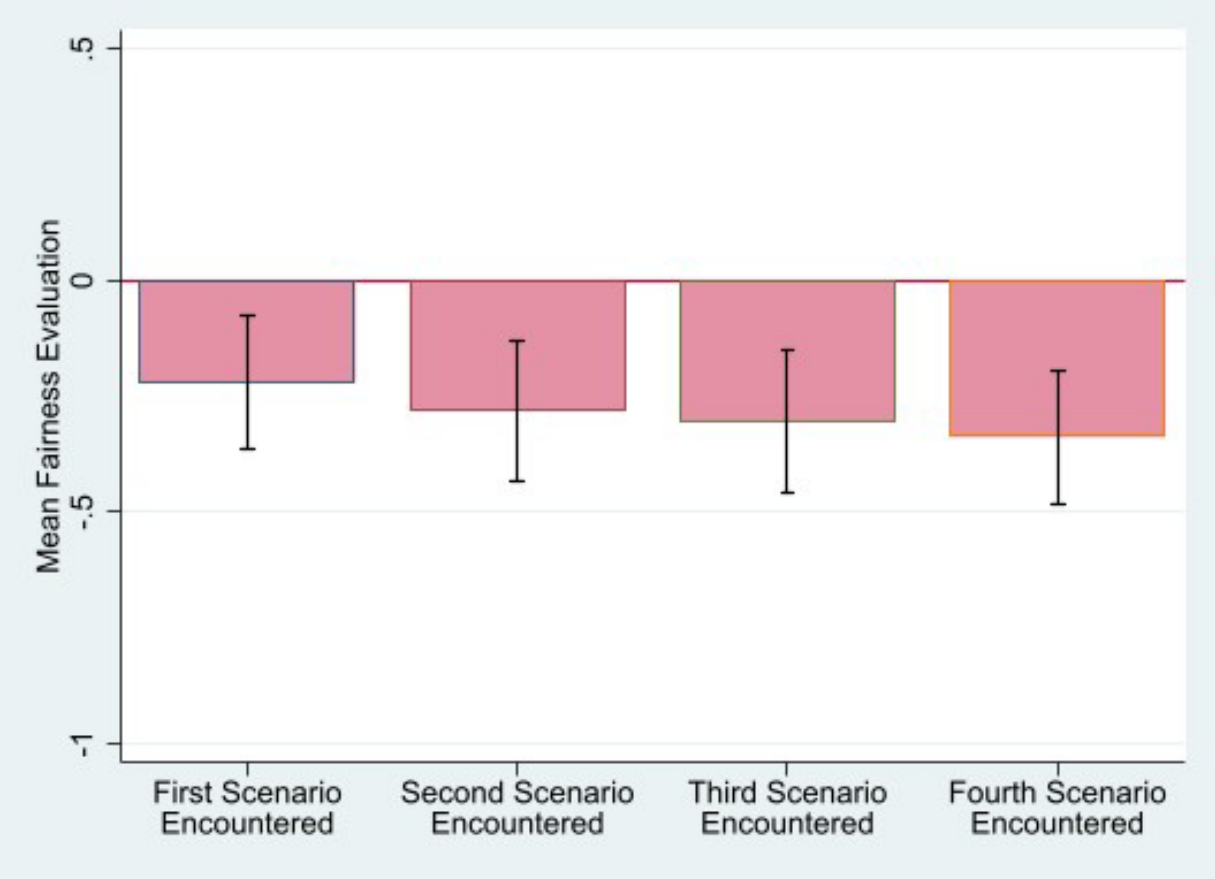
Note: This figure illustrates trends in Google searches for keywords related to three topics: "Black Lives Matter (blm), racism, and discrimination. The vertical axis represents search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity 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 A3.1.1 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

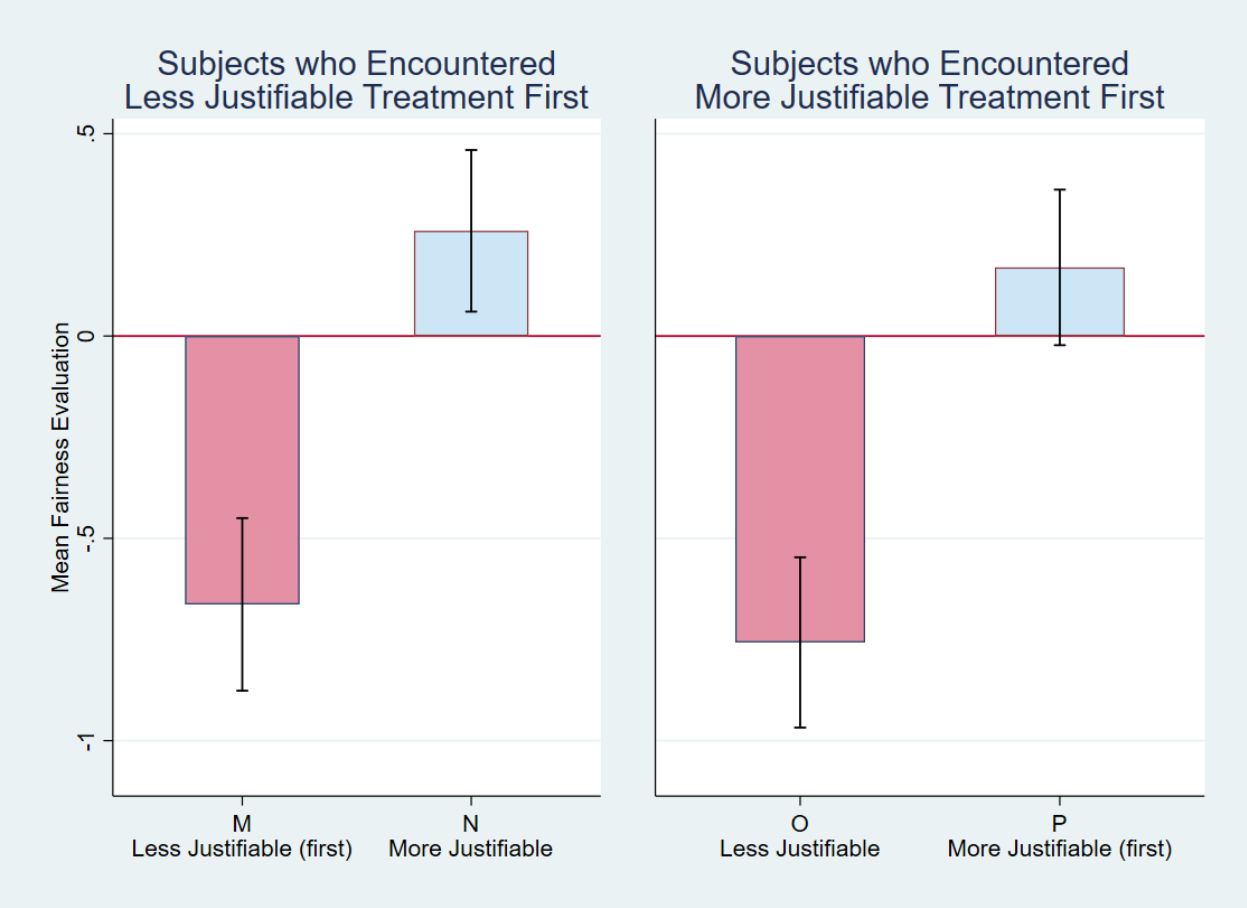
A3.3 Order Effects for the *Less* versus *More* Justifiable Treatments

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.

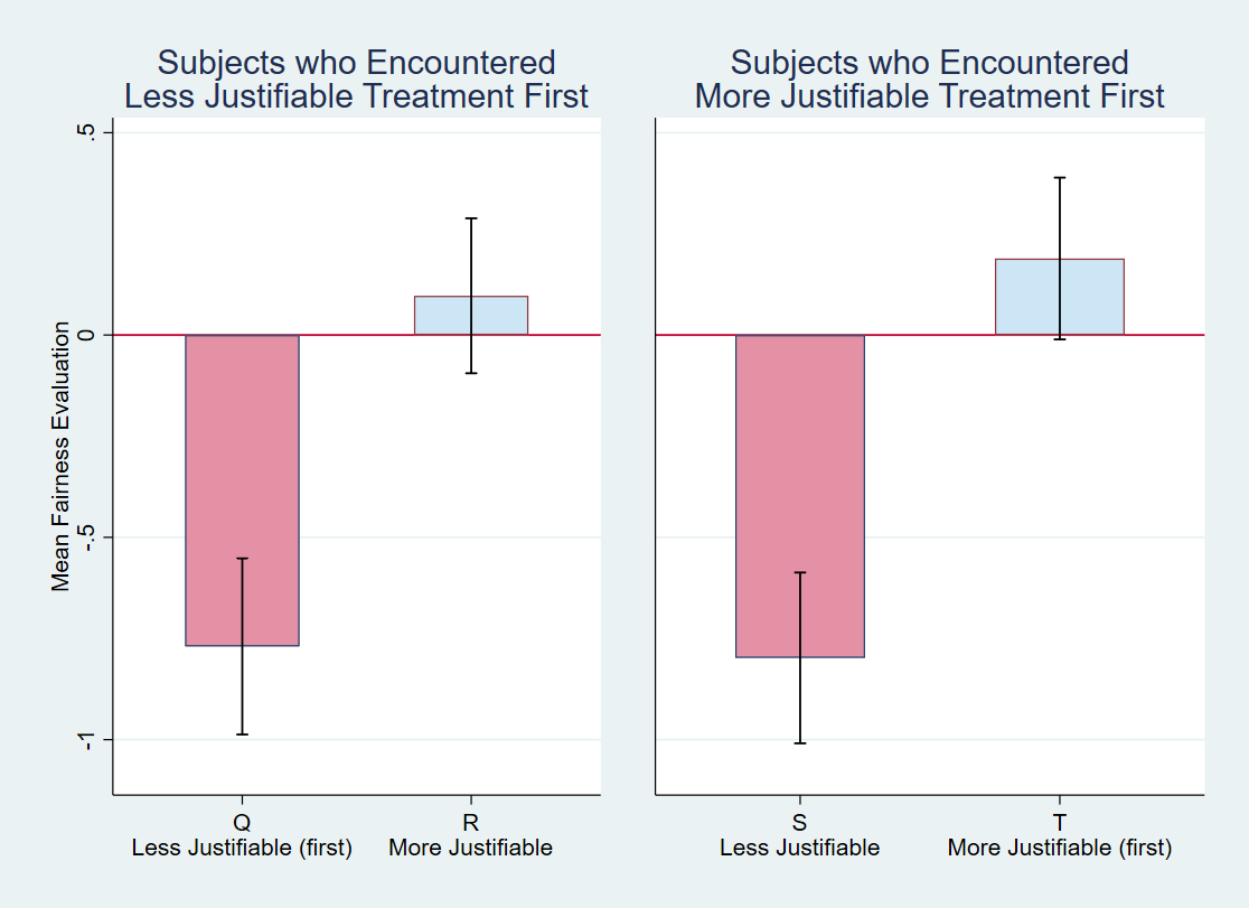
- M vs. N = 0.000
- O vs. P = 0.000
- M vs. O = 0.537
- N vs. P = 0.521

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.

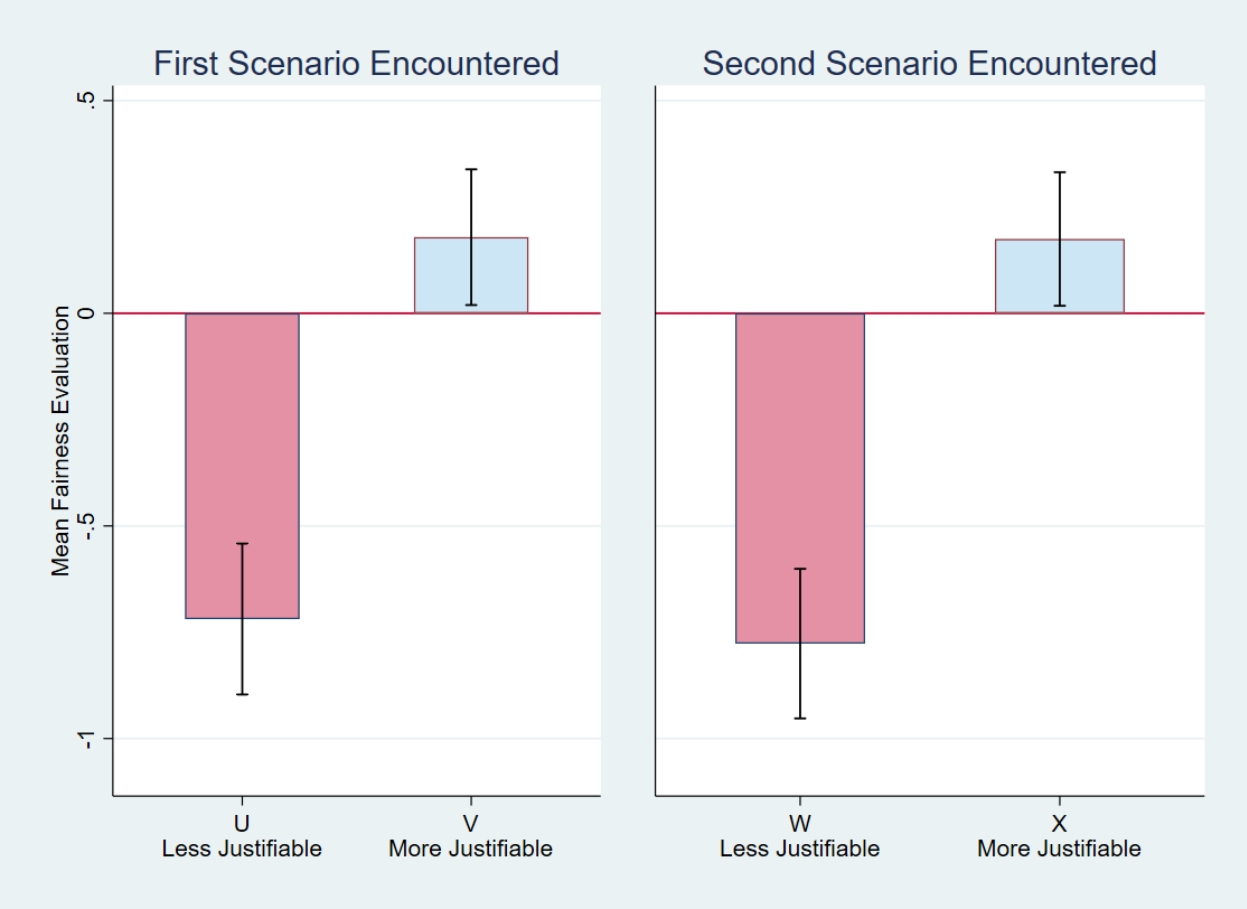
- Q vs. R = 0.000
- R vs. S = 0.000
- Q vs. S = 0.854
- R vs. T = 0.513

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.

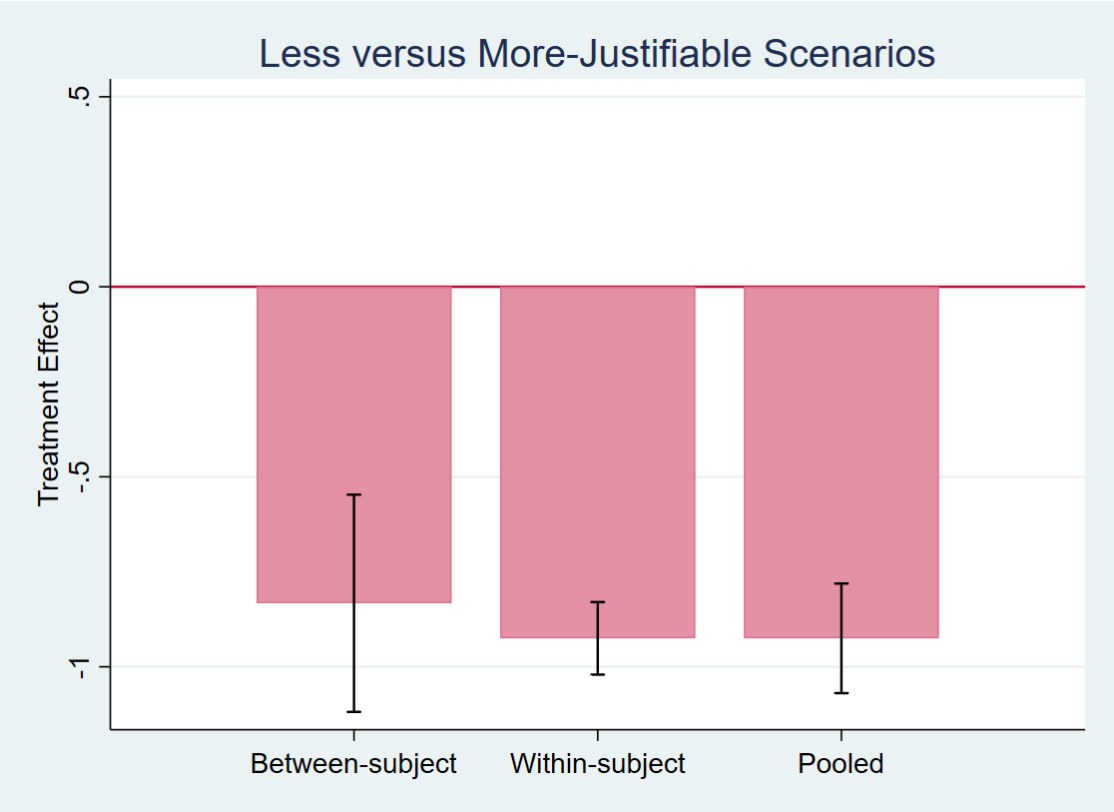
- U vs. V = 0.000
- W vs. X = 0.000
- U vs. W = 0.610
- V vs. X = 0.967

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.

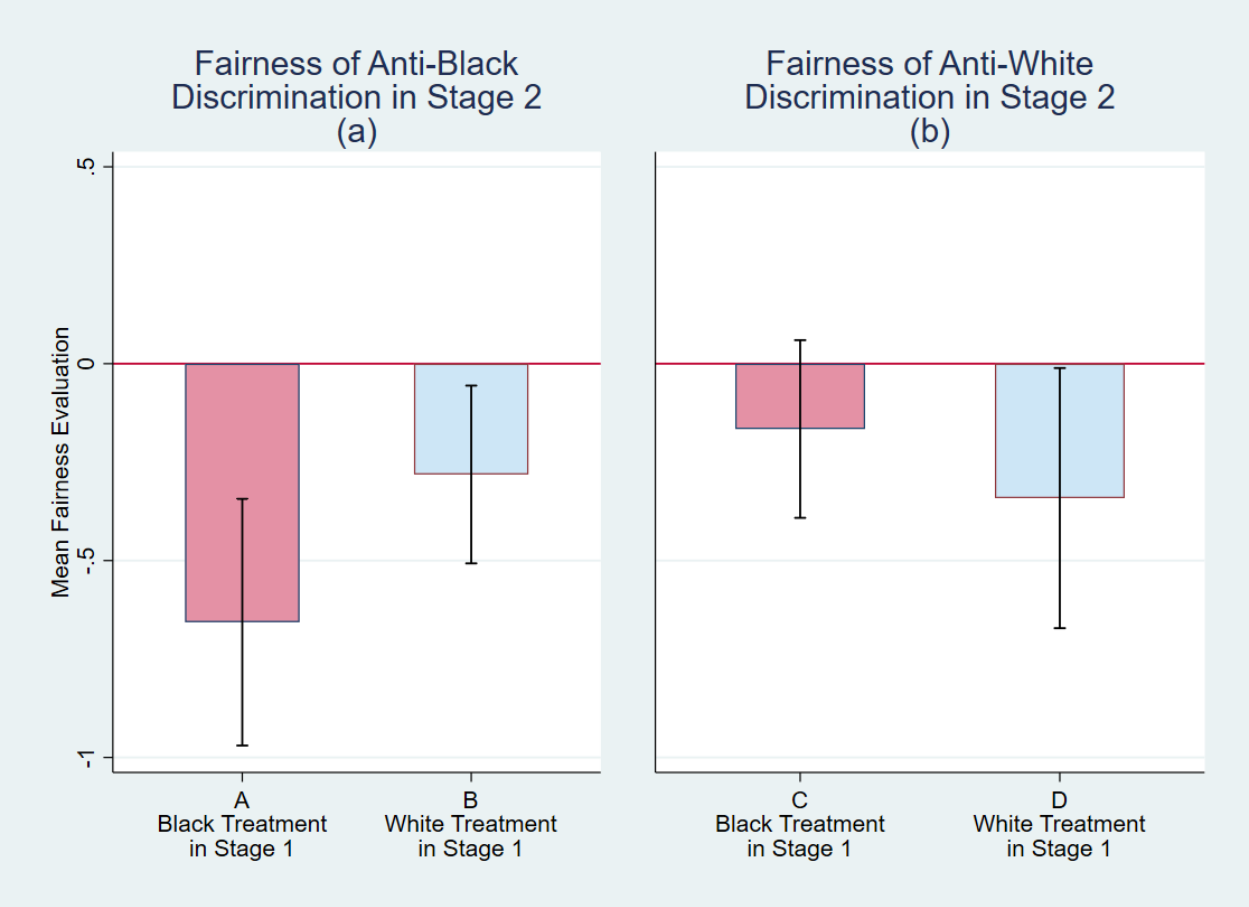
A3.4 Order Effects for the Race Treatment

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

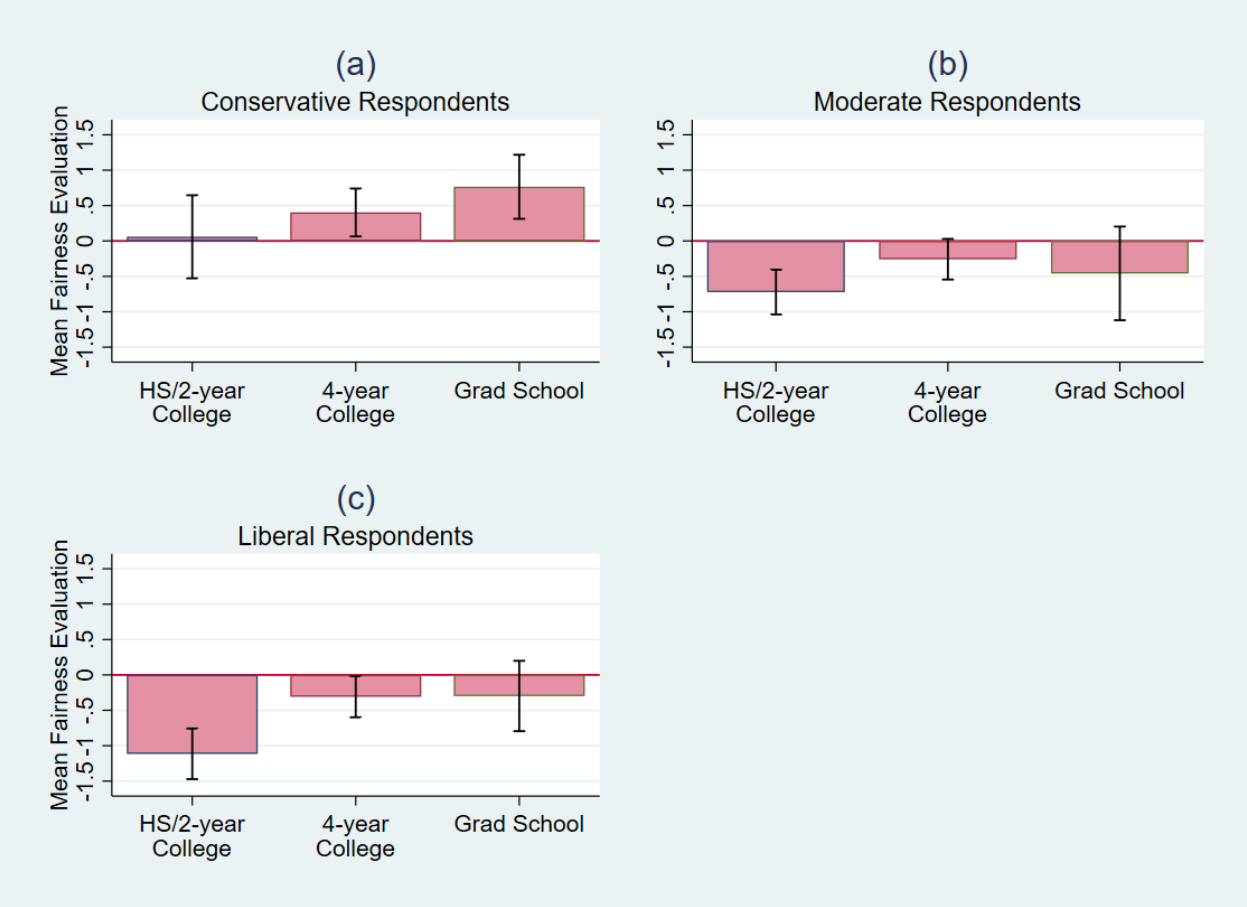
This Section explores the unexpected (to us) positive association between respondents' education and their ratings of the fairness of discriminatory actions. We show, first of all, that the positive association between education and fairness is not an artifact of political differences between the education 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:

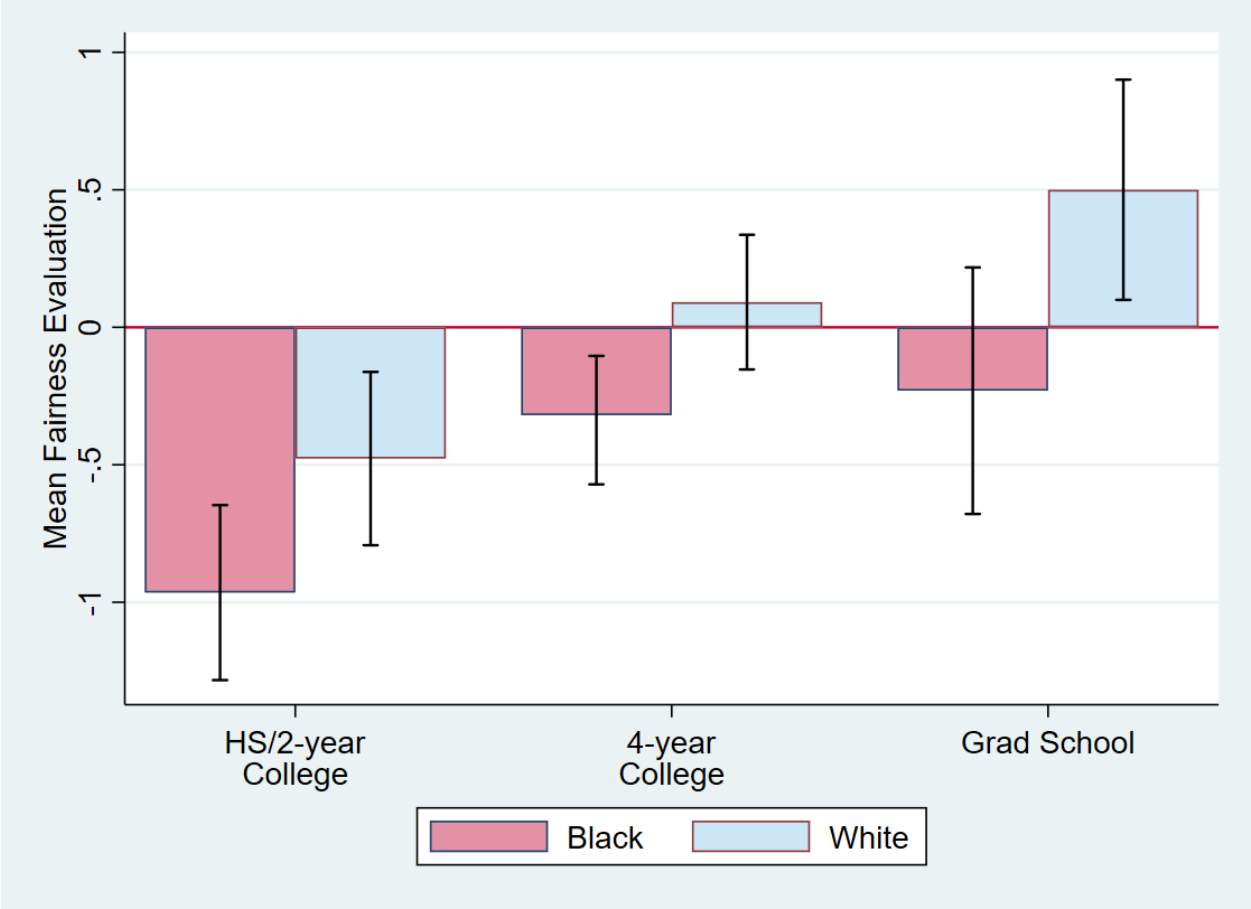


Notes: This figure is based on only Stage 1 observations. The *p*-values below are clustered by respondent.

- 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.

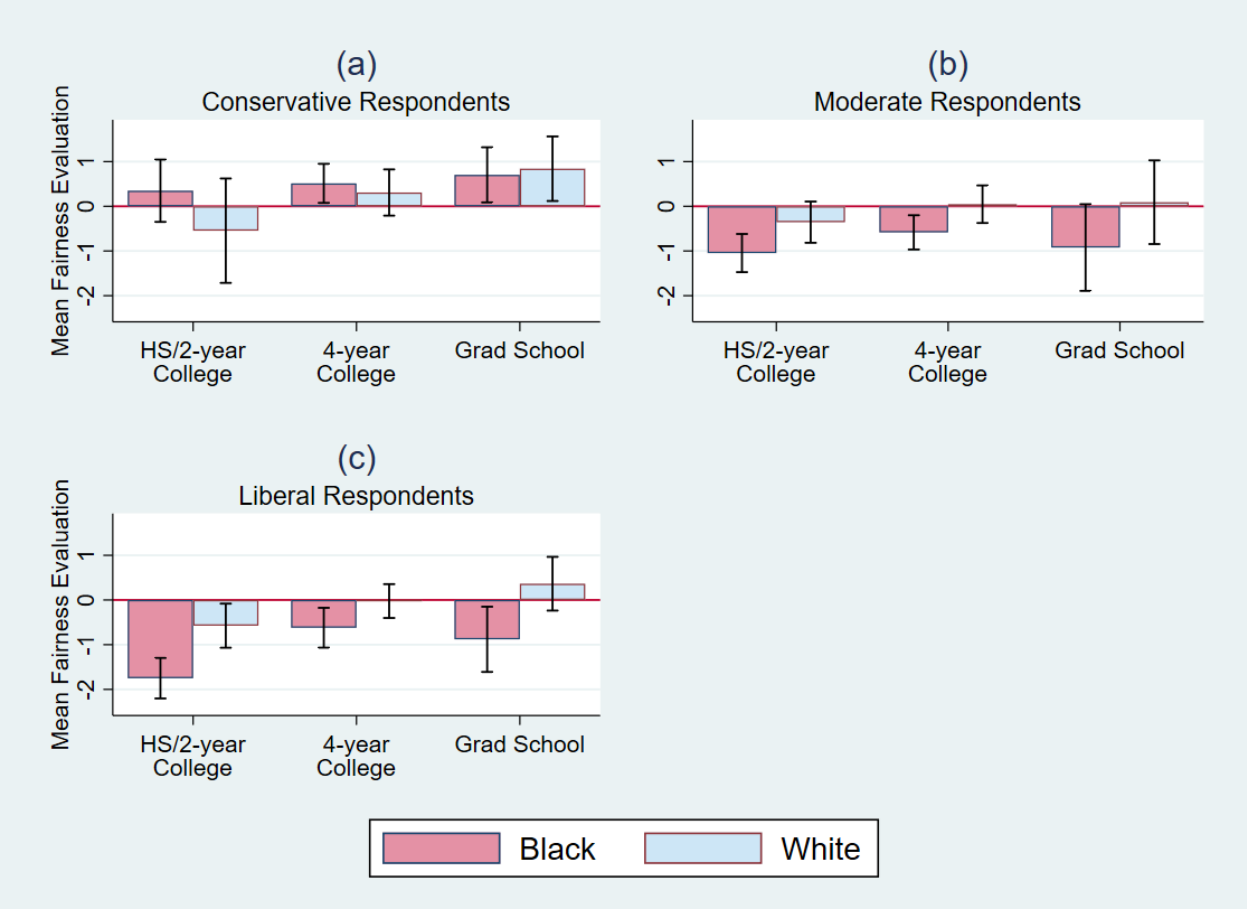


Notes: This figure is based on only Stage 1 observations. The *p*-values below are clustered by respondent.

- 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

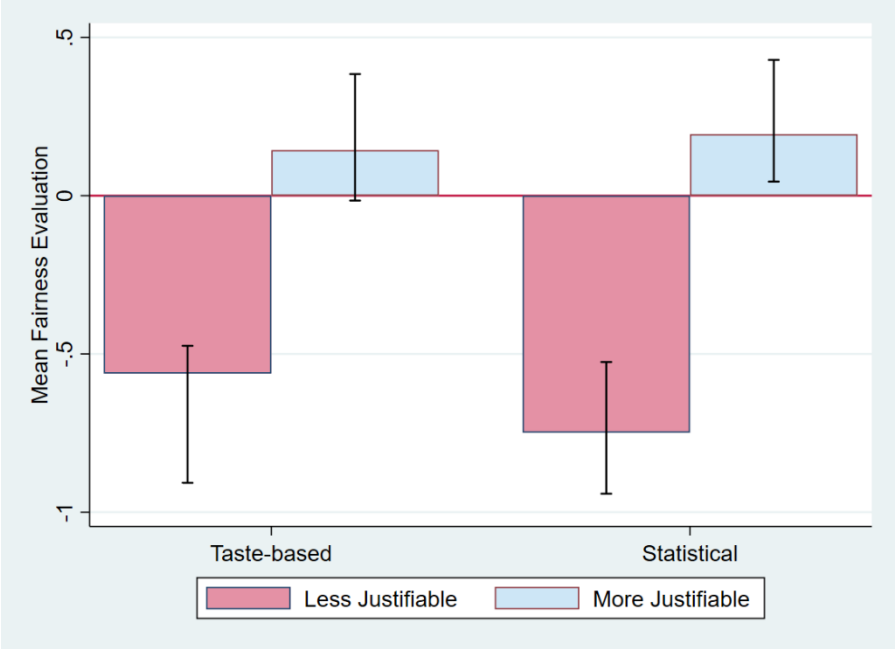
A5.1: Replicating Figures 2 and 3 Using First Scenarios Only

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)

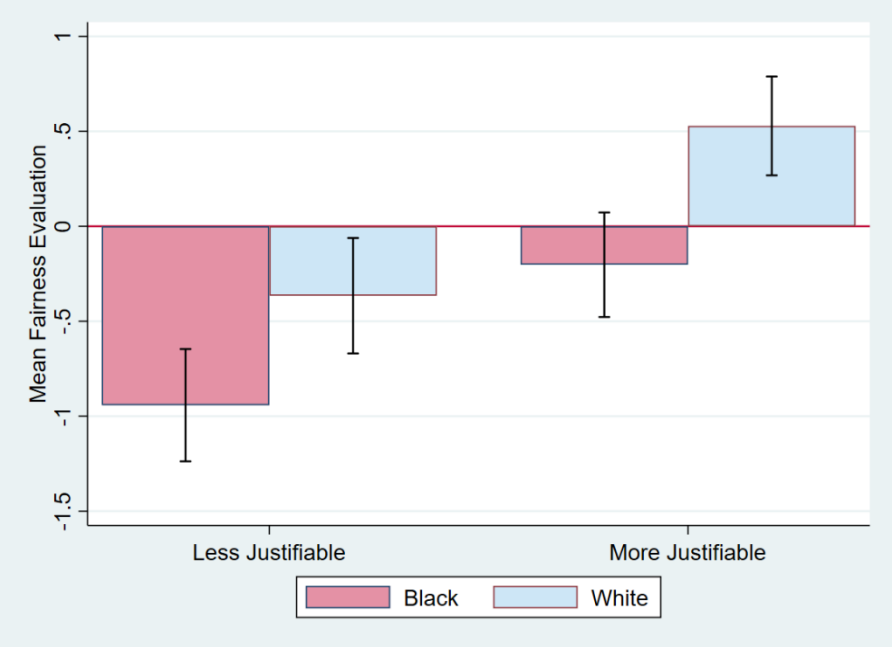


Notes: This figure replicates Figure 2 using only observations from the first scenario encountered by respondents in Stage One. Therefore, the *p*-values displayed below are not clustered.

- 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)



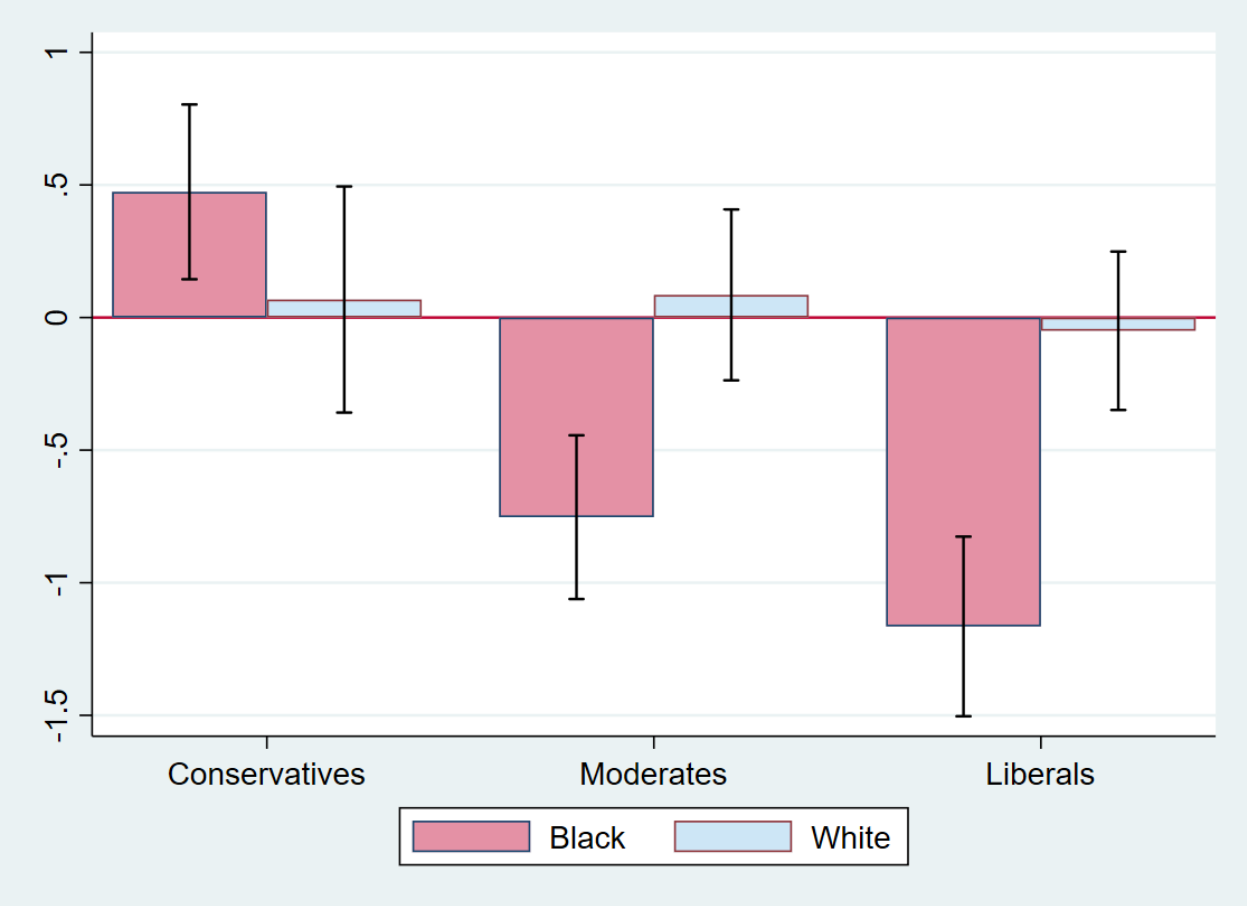
Notes: This figure replicates Figure 3 using only observations from the *first* scenario encountered by respondents in Stage One. Therefore, the *p*-values displayed below are not clustered.

- 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



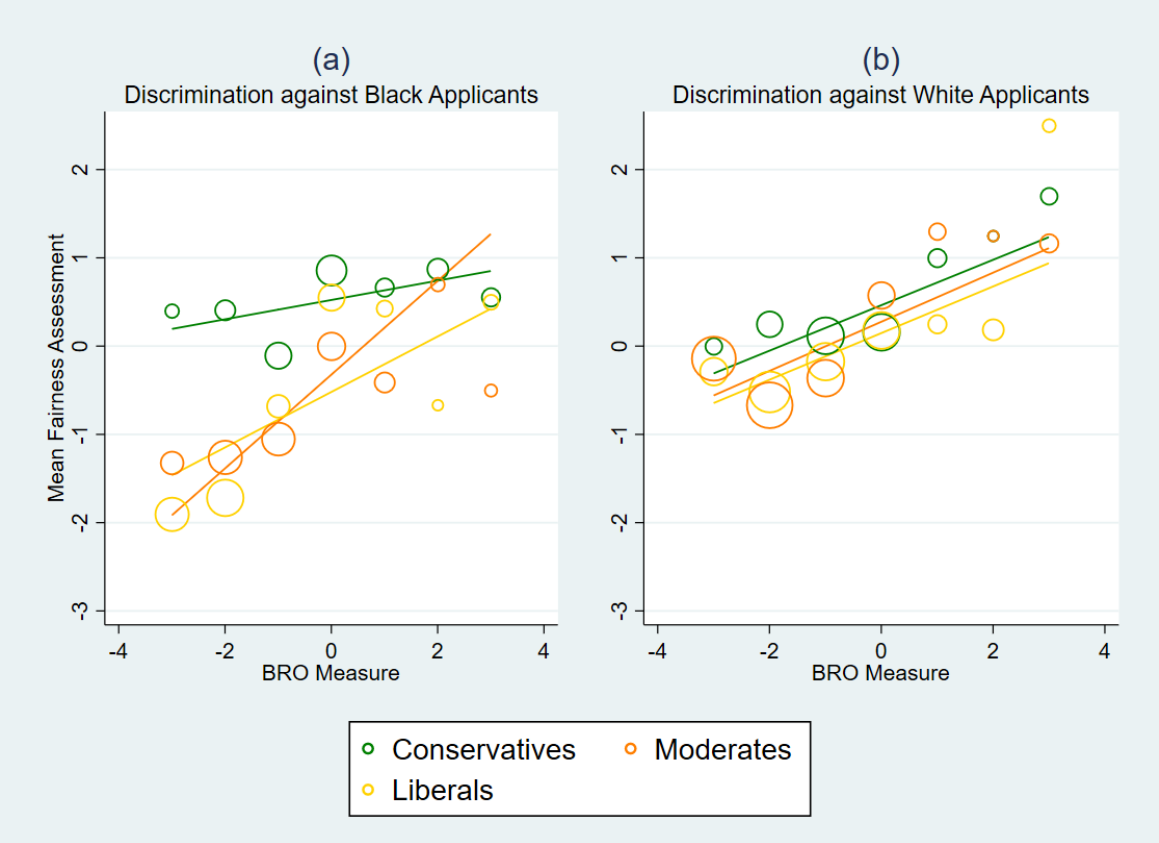
Notes: This figure reproduces Figure 6, but it only reflects the fairness evaluations of White respondents. The p-values below are clustered by respondent.

- 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.[52]

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

[52] 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."

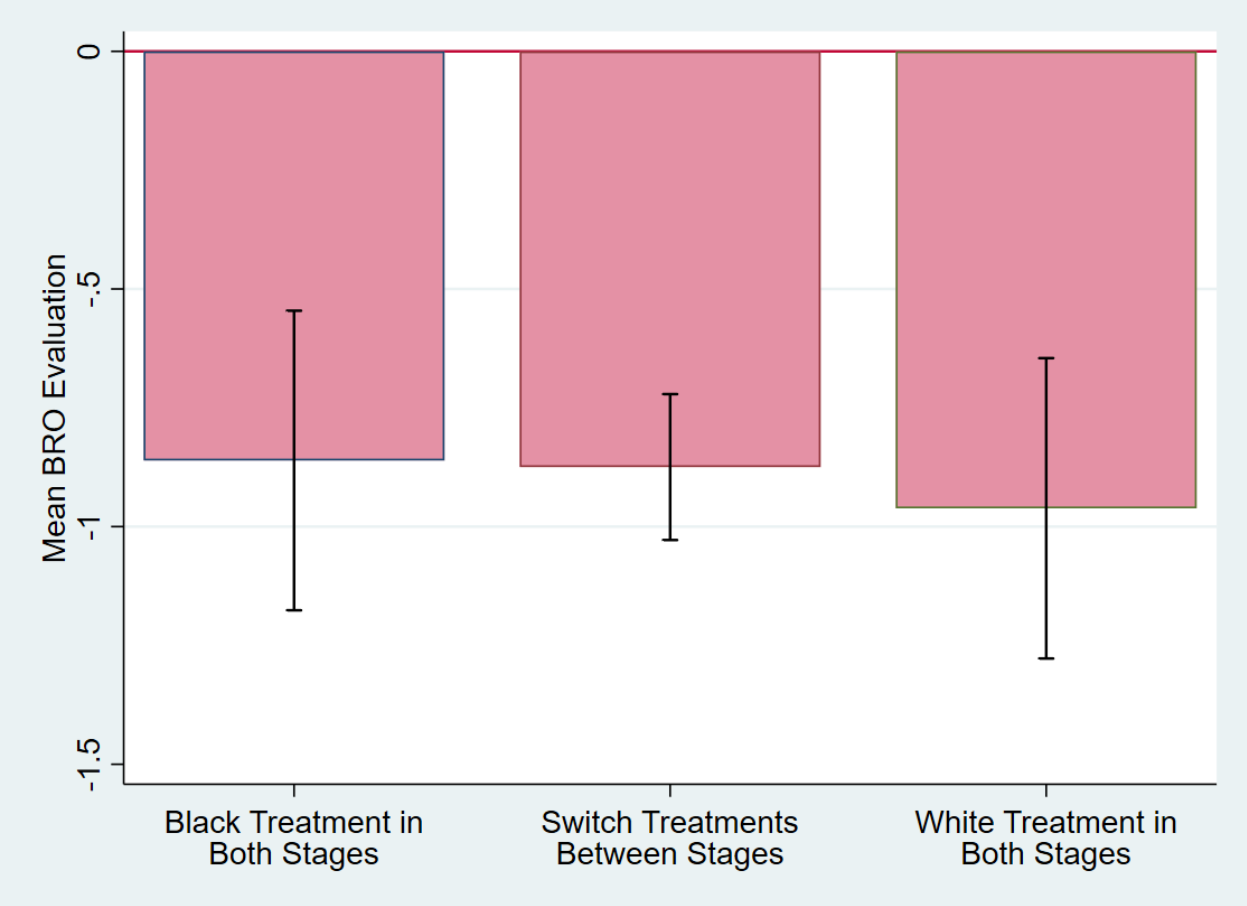
Note: All of these statements (7/7) were for statistical discrimination scenarios.

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*:*

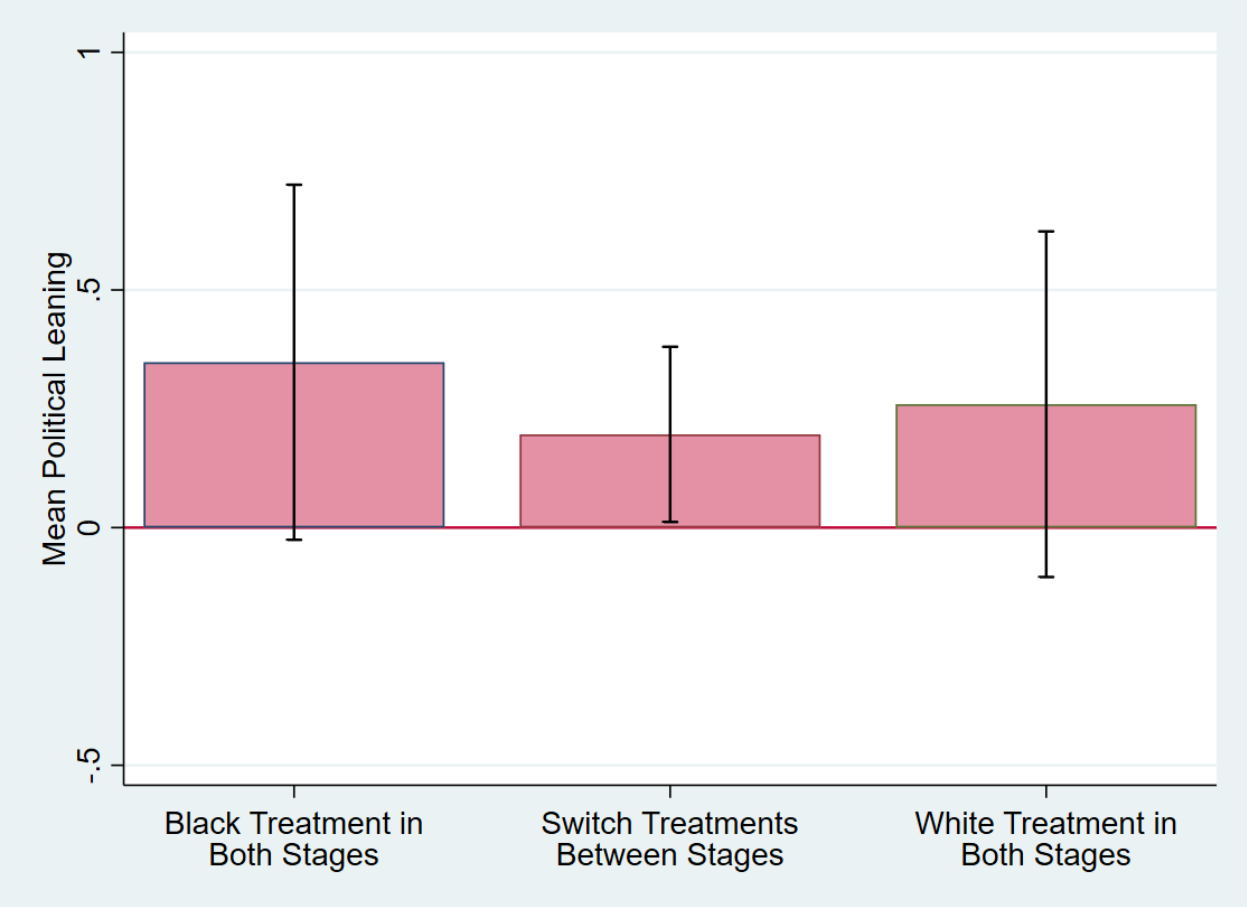
BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more).

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).

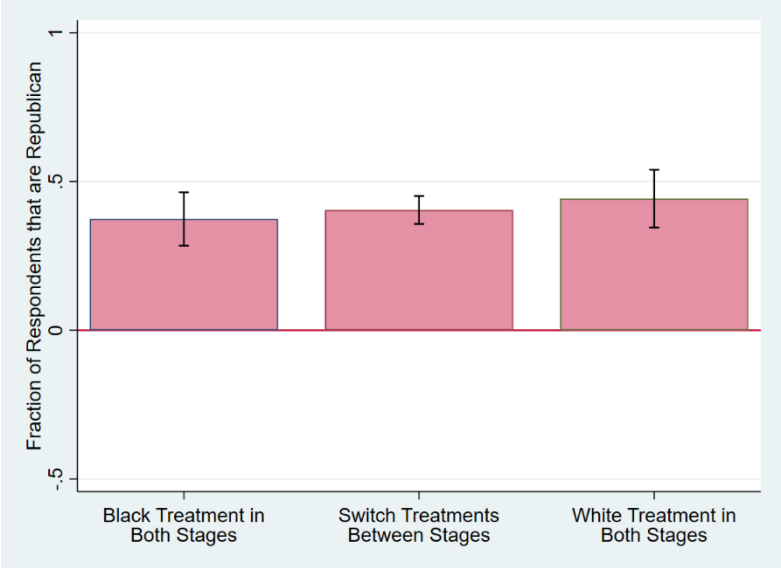
Notes*:* The *p*-values below are clustered by respondent.

- 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

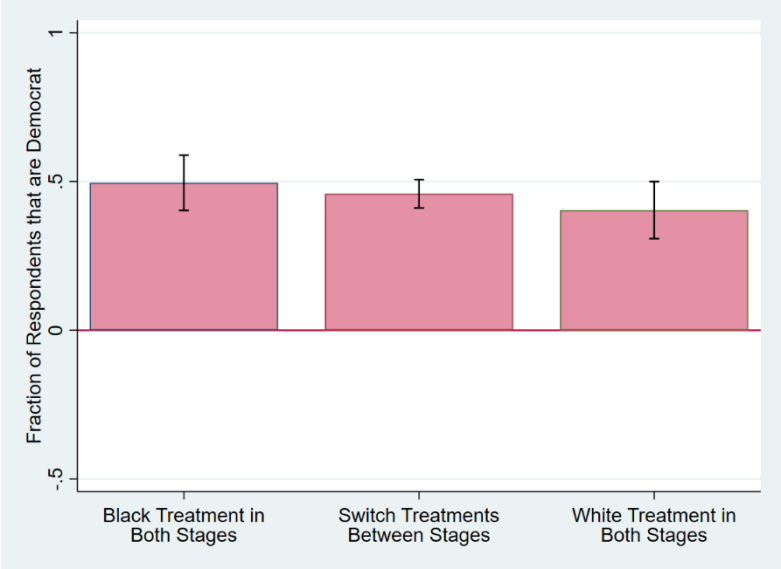
If the respondents modify their reported political leanings to cater to the (inferred) political preferences of the experimenters, we should see a monotonic decrease (shift from liberal towards conservative) from left to right. Such a decrease is not present.

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)

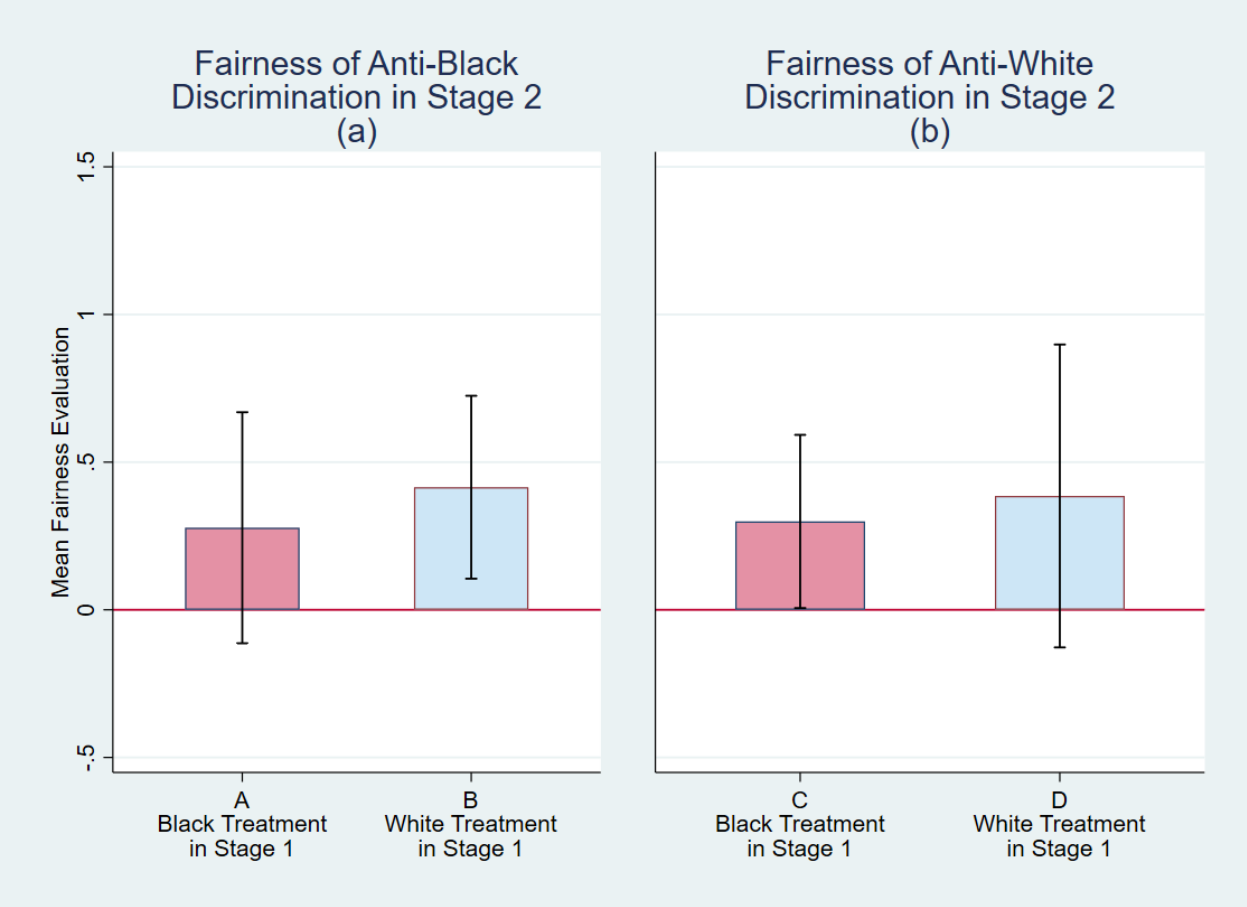


Figure A8.1.1 shows that Group 1's Stage 2 fairness ratings *do not* depend on the discriminatee race they encountered in Stage 1:

***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)

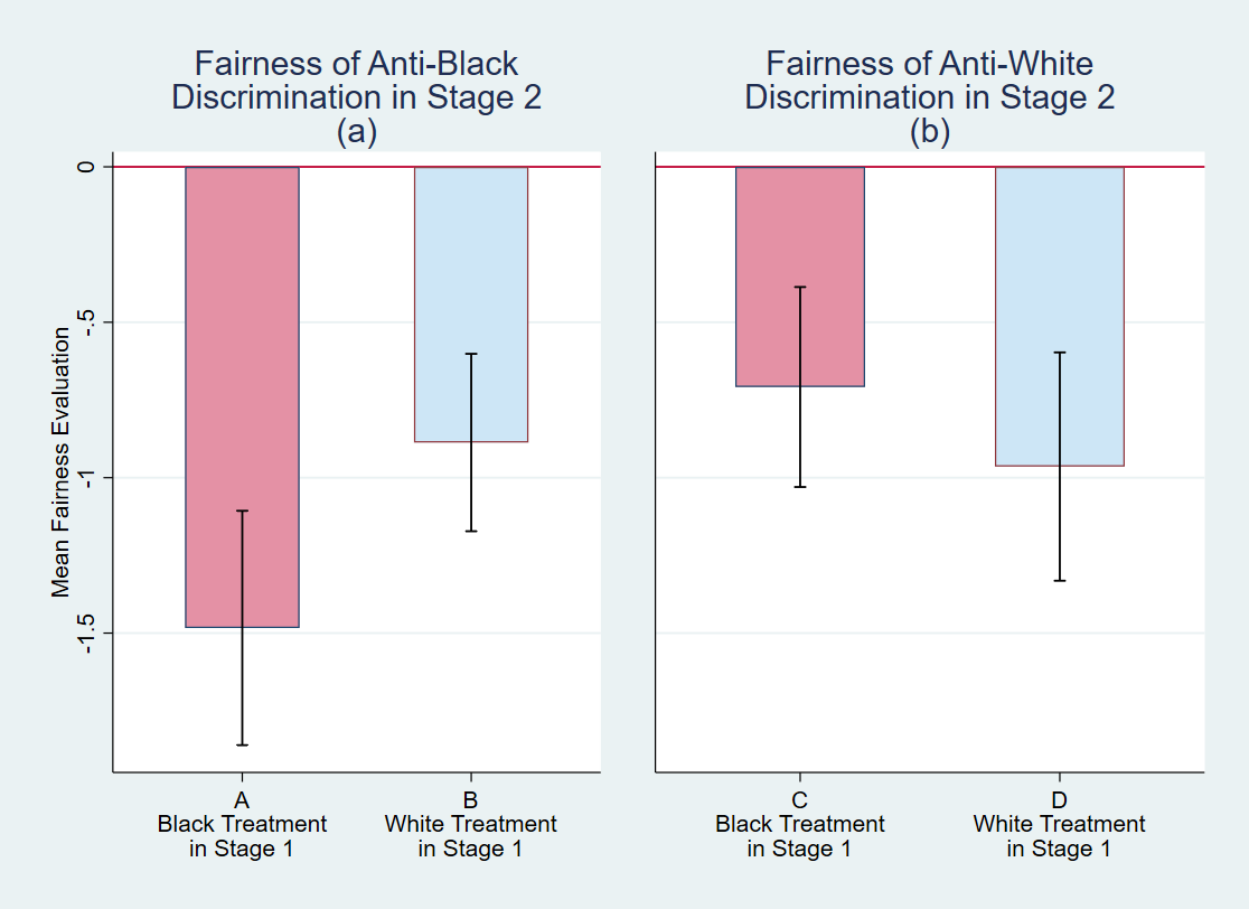


Figure A8.1.2 shows that Group 2's Stage 2 fairness ratings *do* depend on the discriminatee race they encountered in Stage 1. Specifically, respondents who encountered the Black treatment in Stage 2 rated it as much less fair if they encountered a Black discriminatee in Stage 1 than if they encountered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

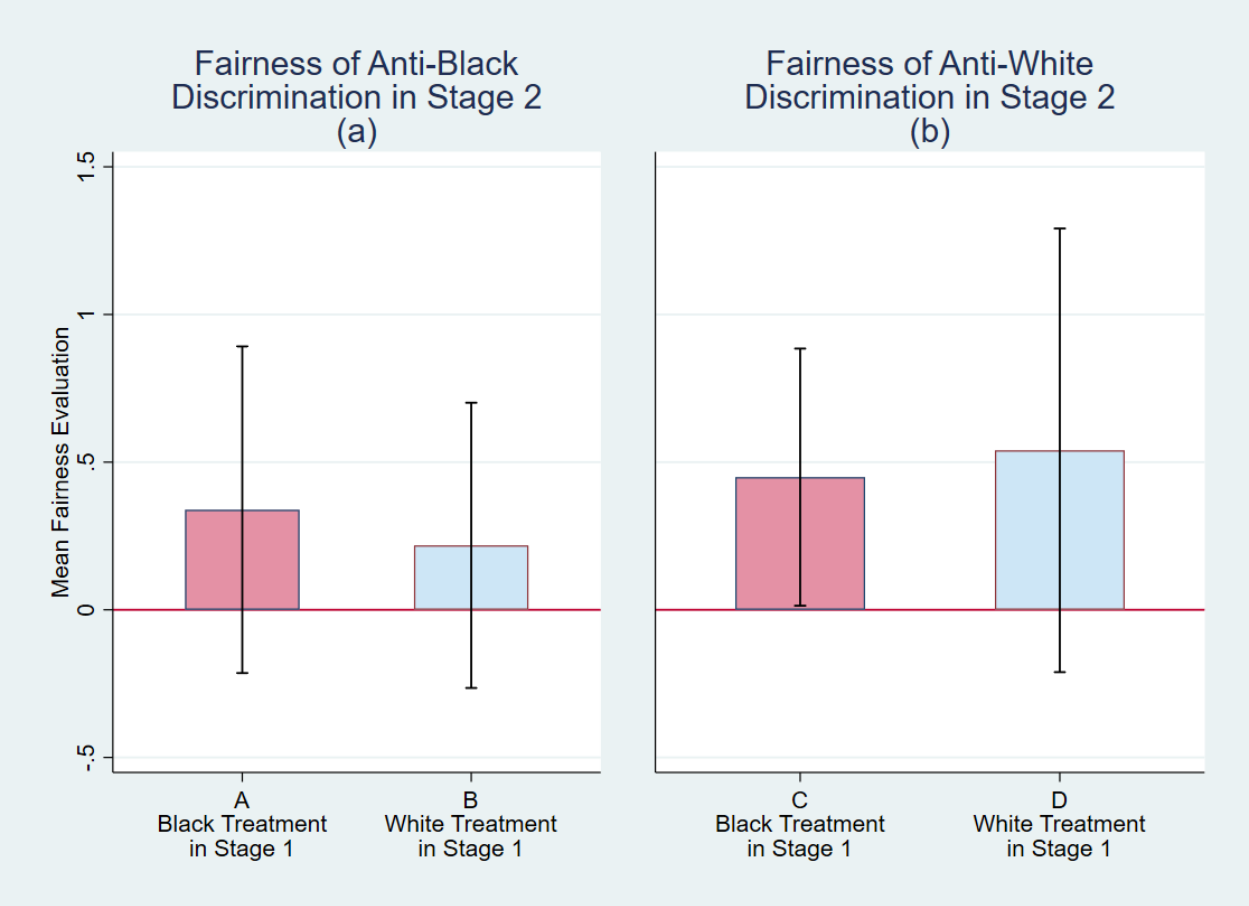


Figure A8.2.1 shows that conservative respondents' Stage 2 fairness ratings *do not* depend on the discriminatee race they encountered in Stage 1:

***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

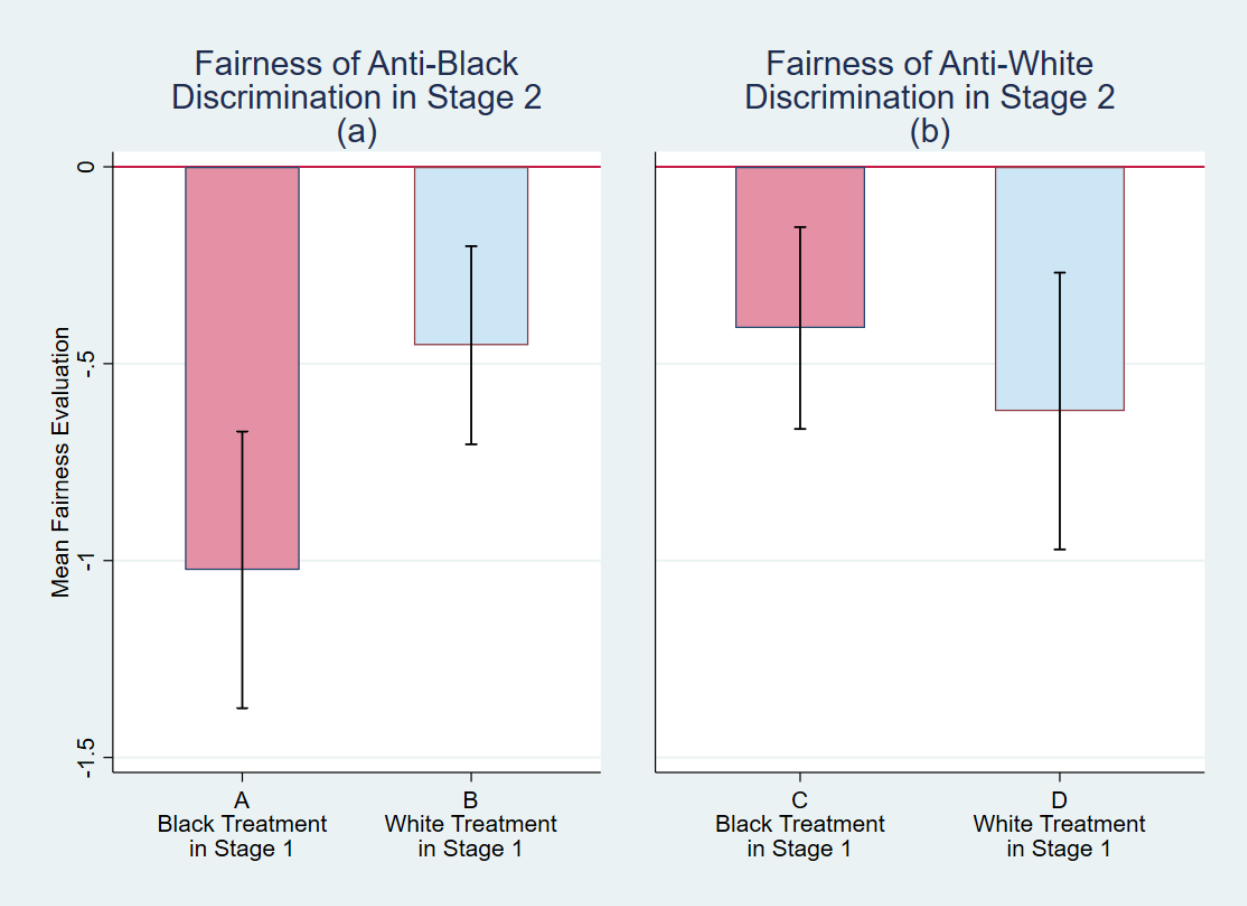


Figure A8.2.2 shows that moderate and liberal respondents' Stage 2 fairness ratings *do* depend on the discriminatee race they encountered in Stage 1. Specifically, respondents who encountered the Black treatment in Stage 2 rated it as much less fair if they also encountered it in Stage 1 than if they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

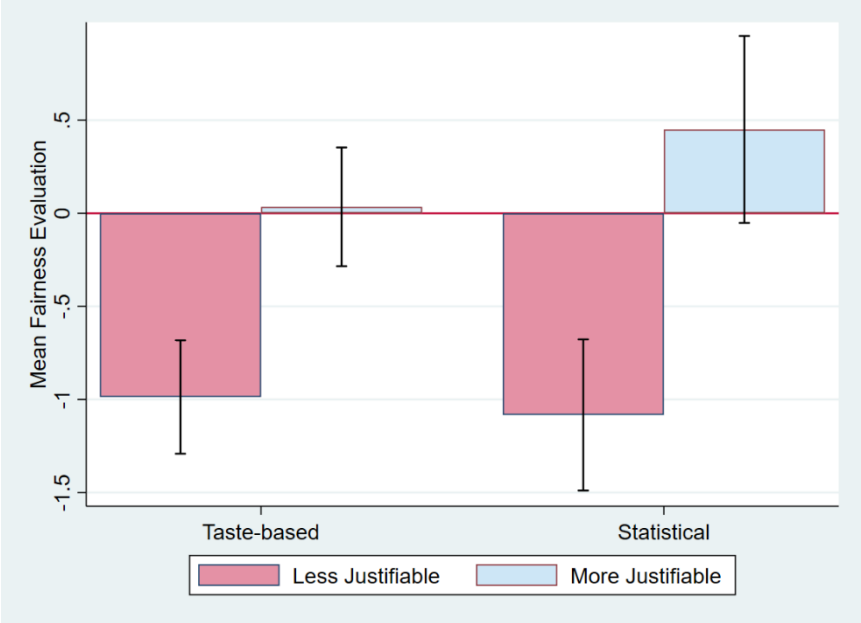
The remaining exhibits in this Appendix replicate Figures 2-8 using these weights. All the main patterns discussed in the paper are also present here, with one small exception: the weak positive association between BRO and the fairness of discrimination among conservative respondents in Figure 8(a) becomes 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)



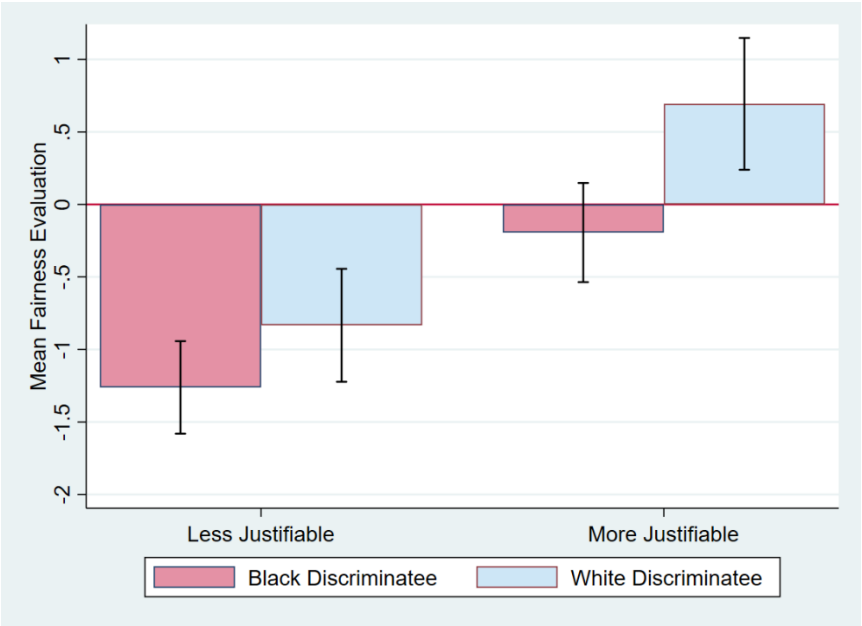
p-values:

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:

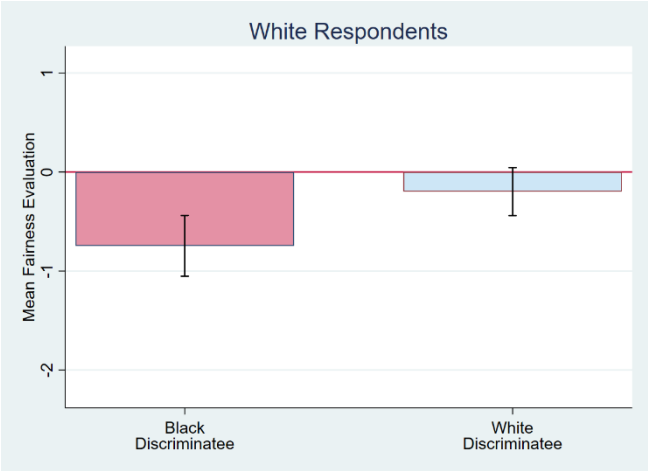
Black versus White Treatment:
Overall: *p*=.003
Within Less-Justifiable: *p*=.095
Within More-Justifiable: *p*=.002

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

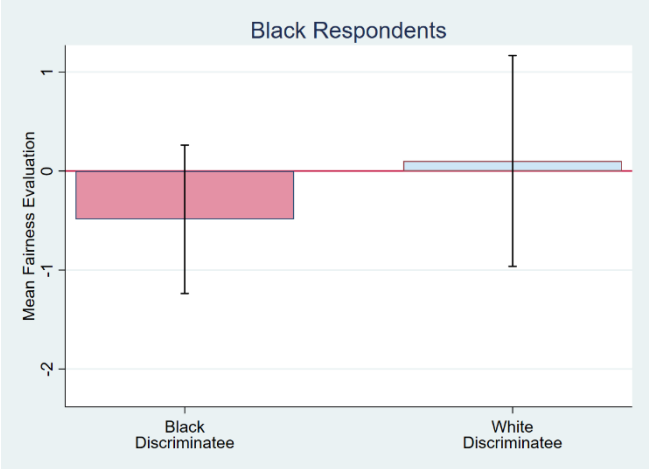
Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 1.068 units less fair. Within White Discriminatees, less-justifiable scenarios are 1.527 units less fair. A test for equality of the Less versus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

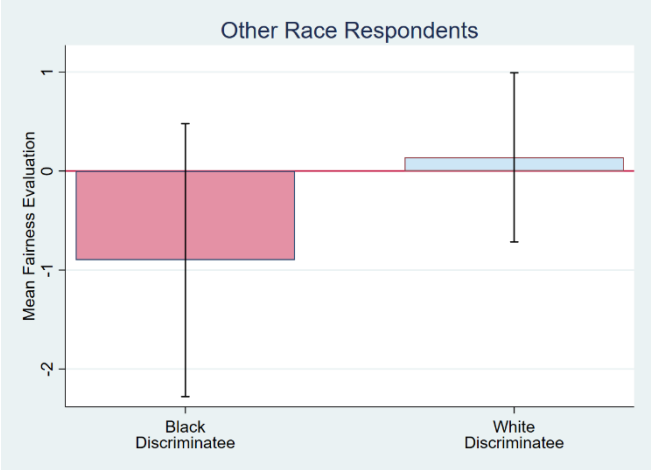
(a)



(b)



(c)

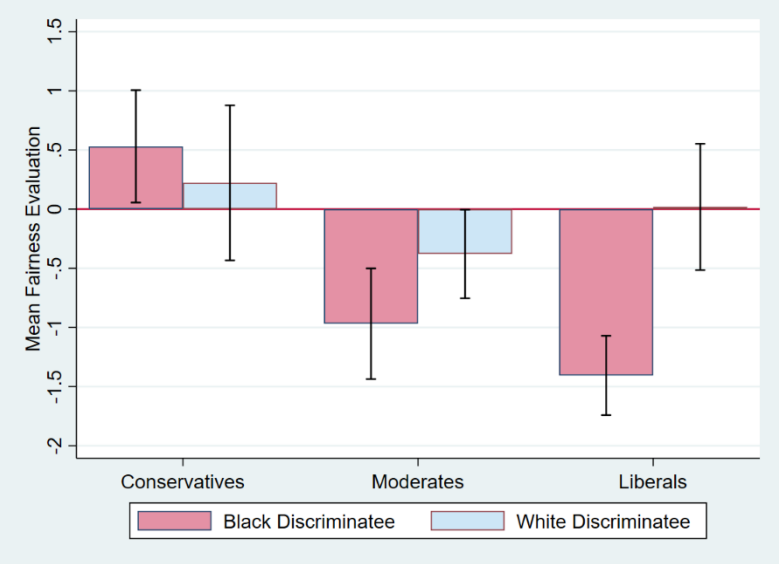


p-values:

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)



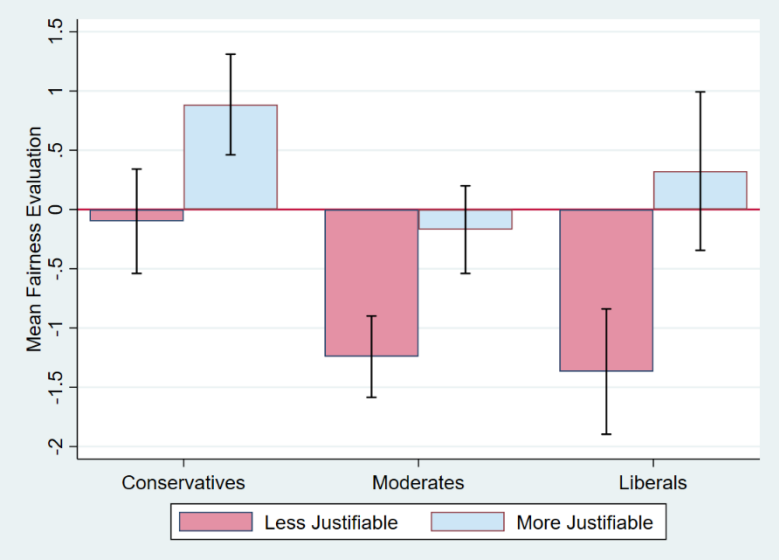
p-values:

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)



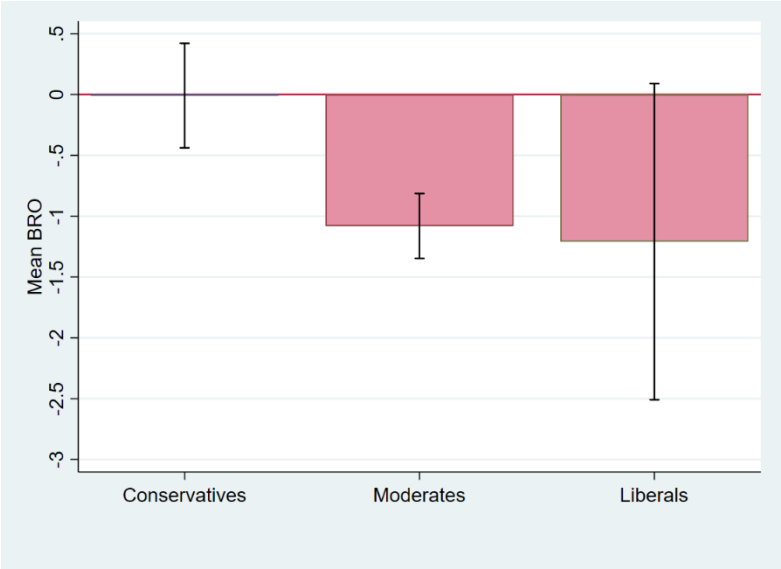
p-values:

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:

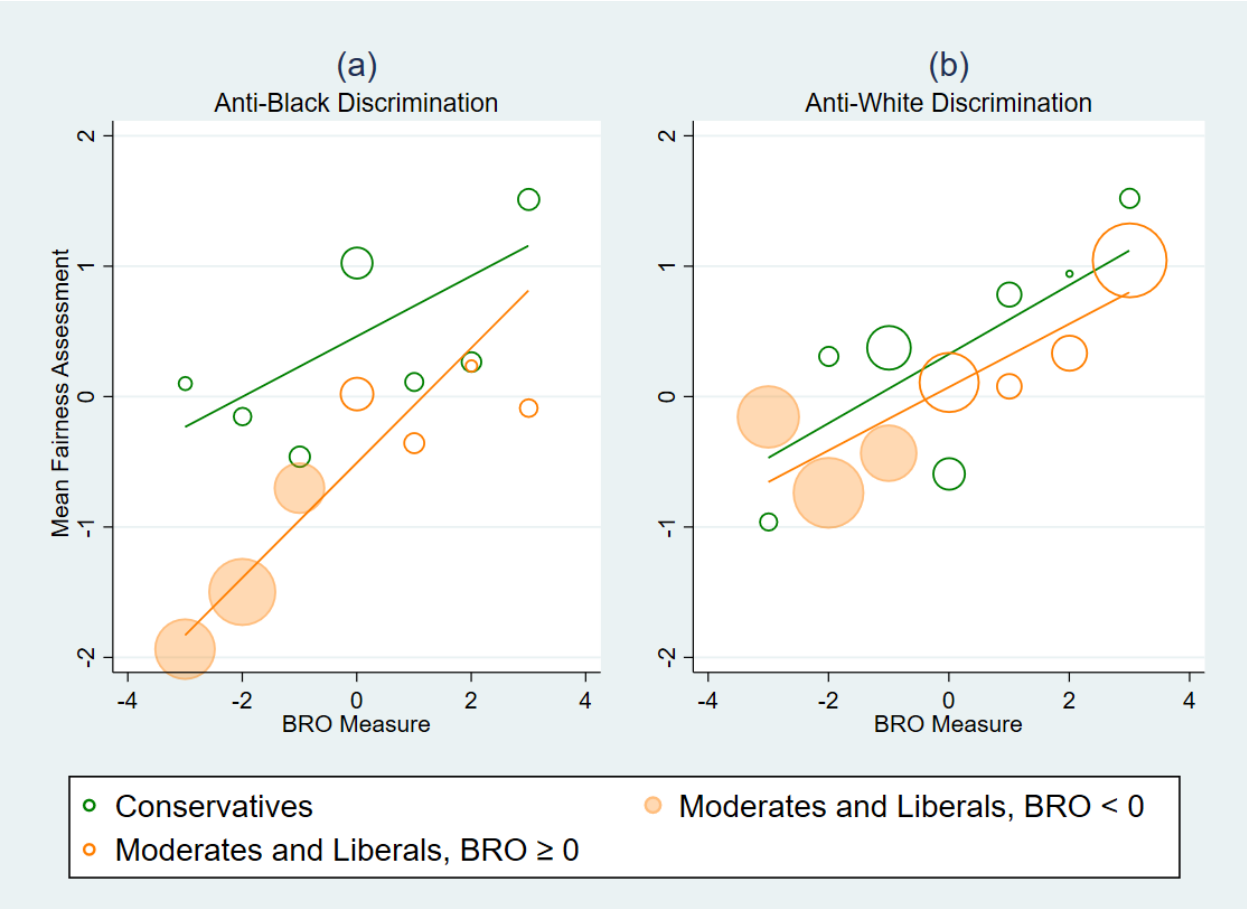
Conservatives vs. Moderates: *p*=.000

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.[53] 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.[54]

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.

[53] 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.

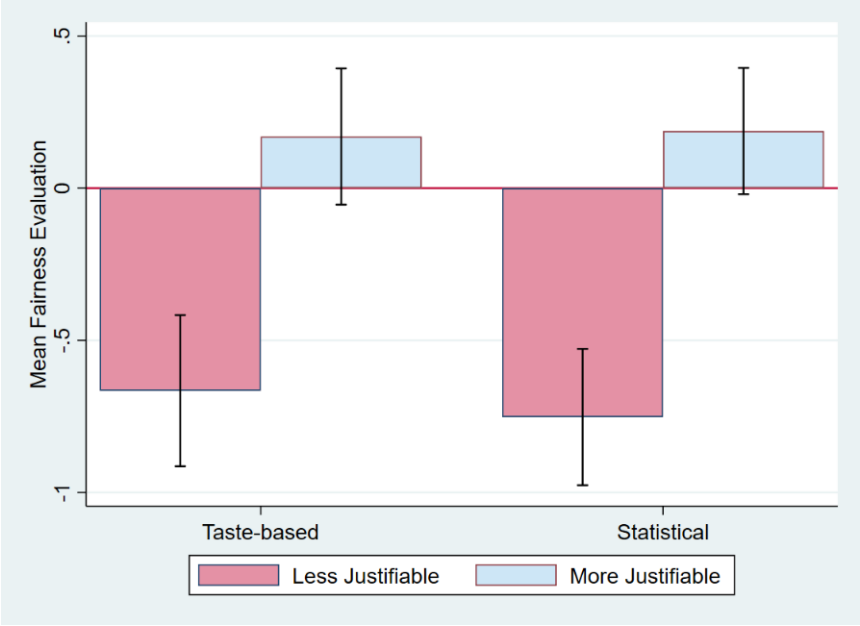
[54] Because of the small size of the MTurk and GSS samples, we did not re-weight our MTurk sample to mimic GSS demographic characteristics; attempts to do this yielded highly extreme and imprecise weights.

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)



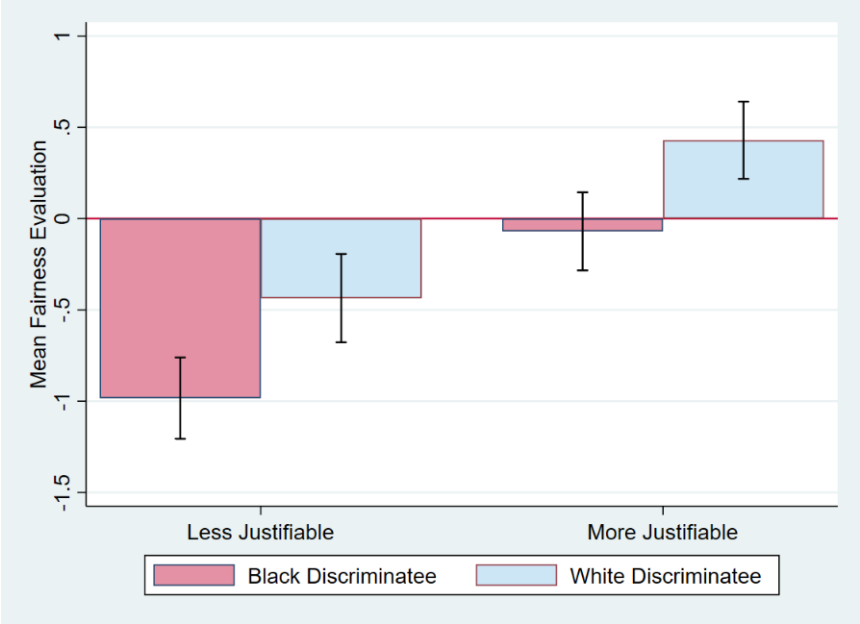
p-values:

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:

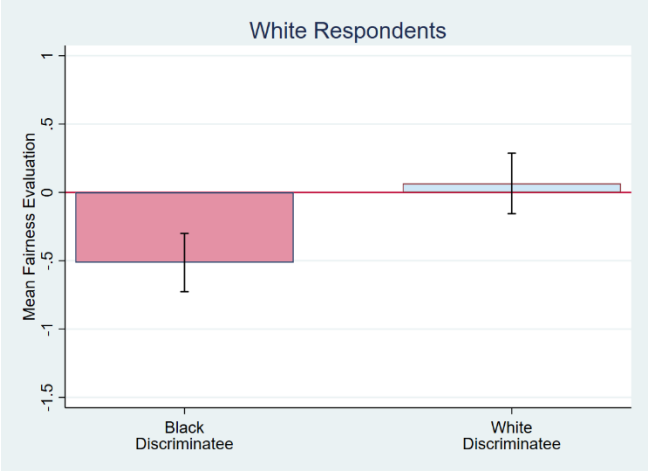
Black versus White Treatment:
Overall: *p*=.000
Within Less-Justifiable: *p*=.001
Within More-Justifiable: *p*=.001

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

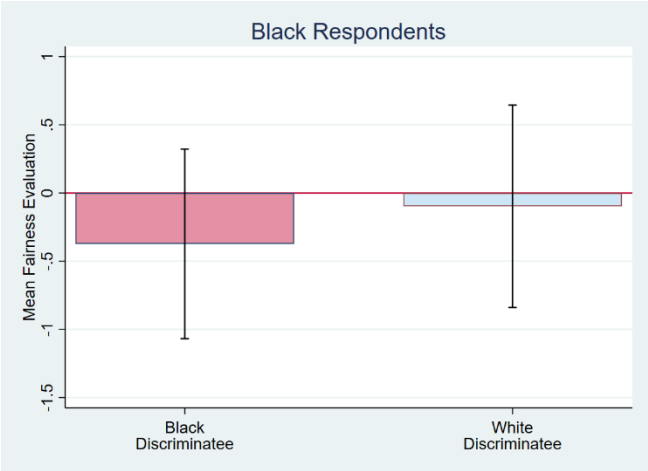
Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.914 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.865 units less fair. A test for equality of the Less versus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

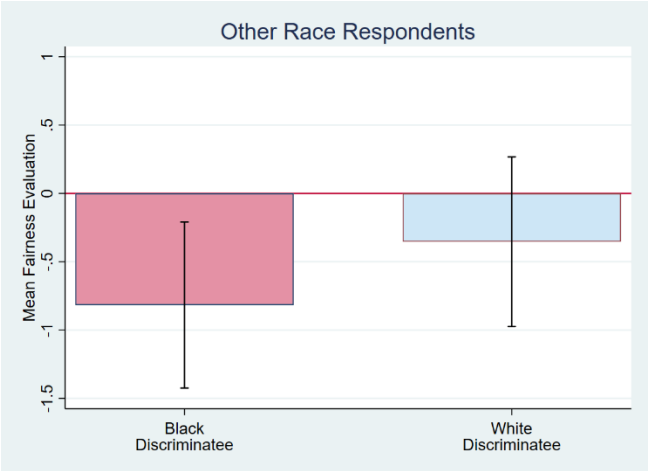
(a)



(b)



(c)



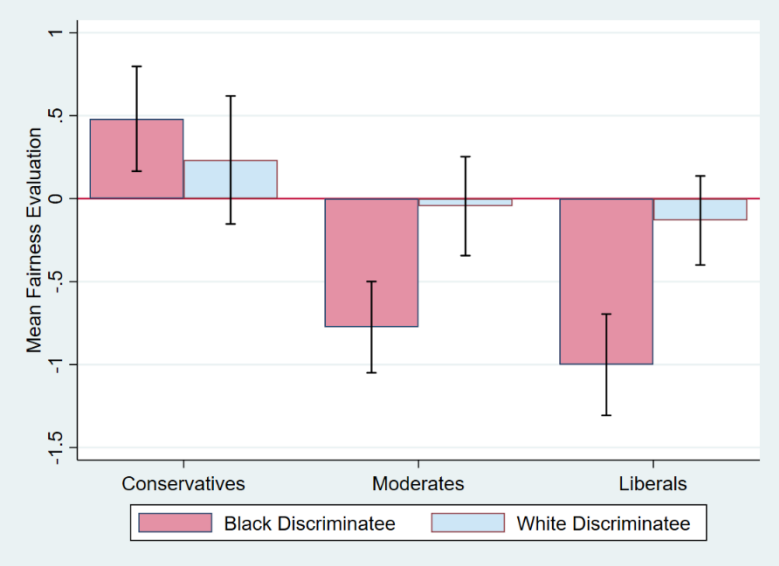
p-values:

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)

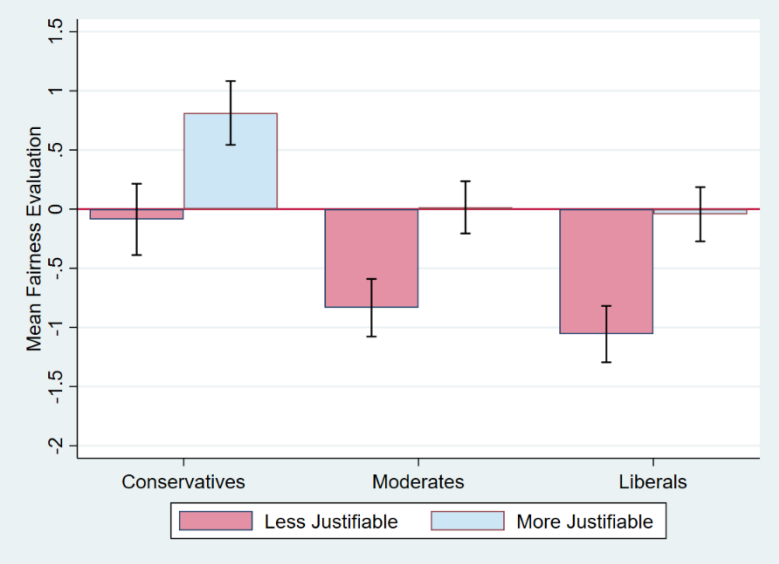


p-values:

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)

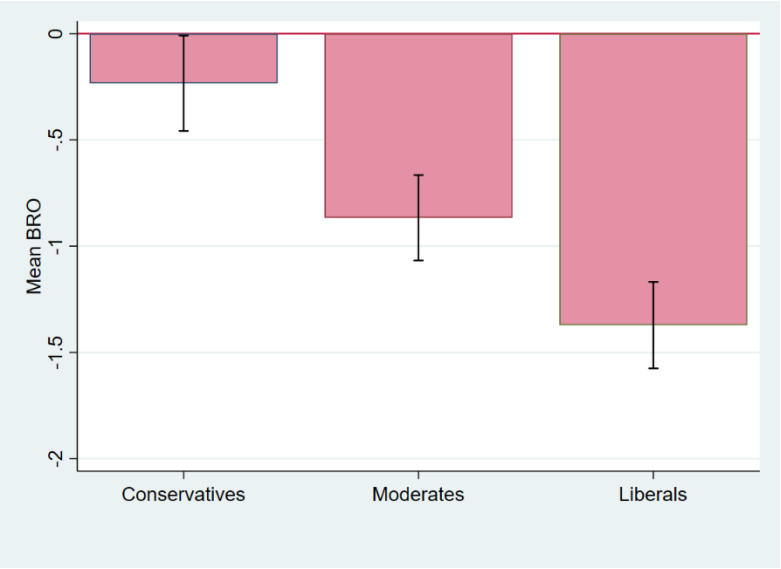


p-values:

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

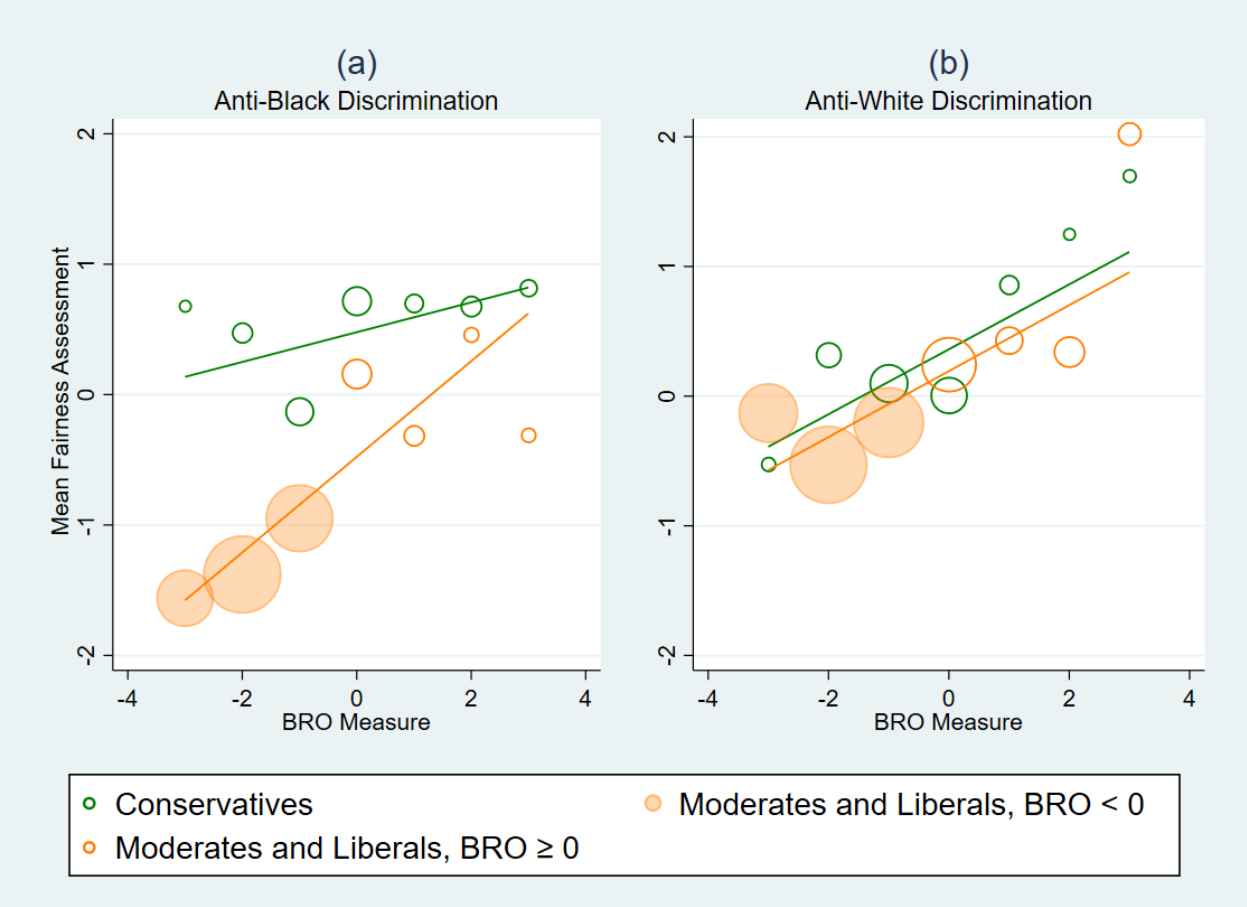
Moderates vs. Liberals: *p*=.001

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).

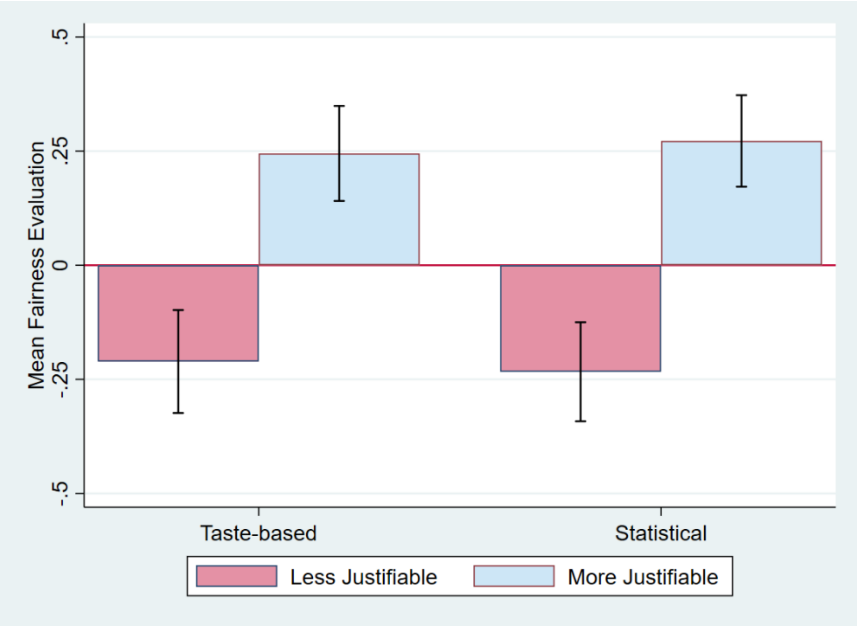
- Panel (a), Discrimination against Black Applicants
  - For Conservatives: slope = 0.114, p = .231
  - For Moderates and Liberals, slope = 0.367, p = .000
- Panel (b), Discrimination against White Applicants
  - For Conservatives: slope = 0.250, p = .096
  - For Moderates and Liberals, slope = 0.254, 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.[55] 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.

[55] 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:

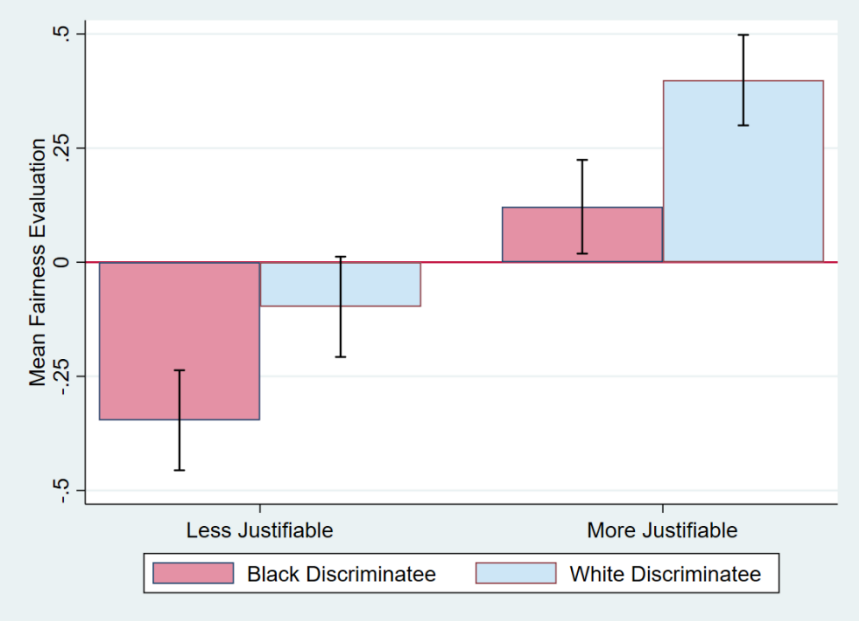
Overall: *p*=.971

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

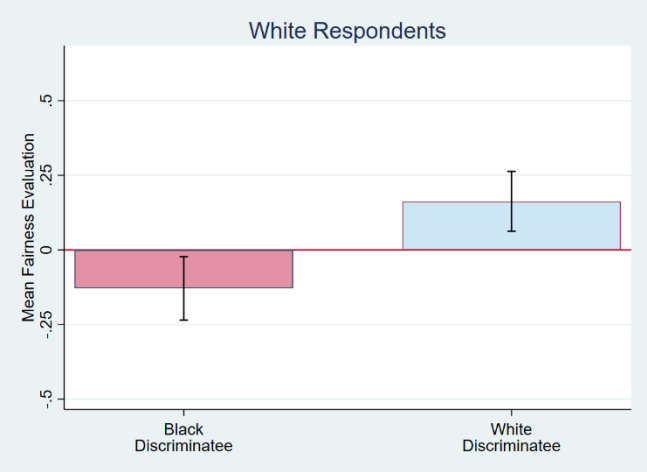
Within Black Discriminatees: *p*=.000

Within White Discriminatees: *p*=.000

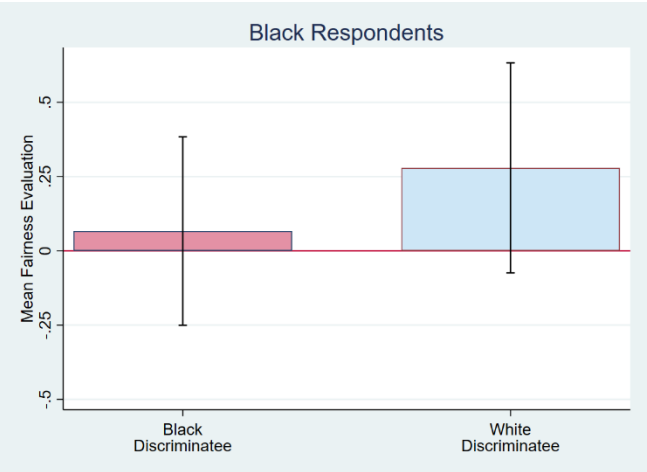
Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.469 standard deviations less fair. Within White Discriminatees, less-justifiable scenarios are 0.495 standard deviations less fair. A test 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)

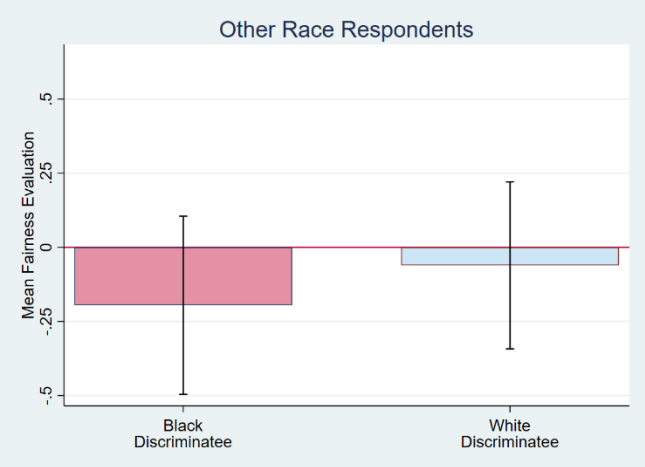
(a)



(b)



(c)



p-values:

Black versus White Treatment:

Overall: *p*=.000

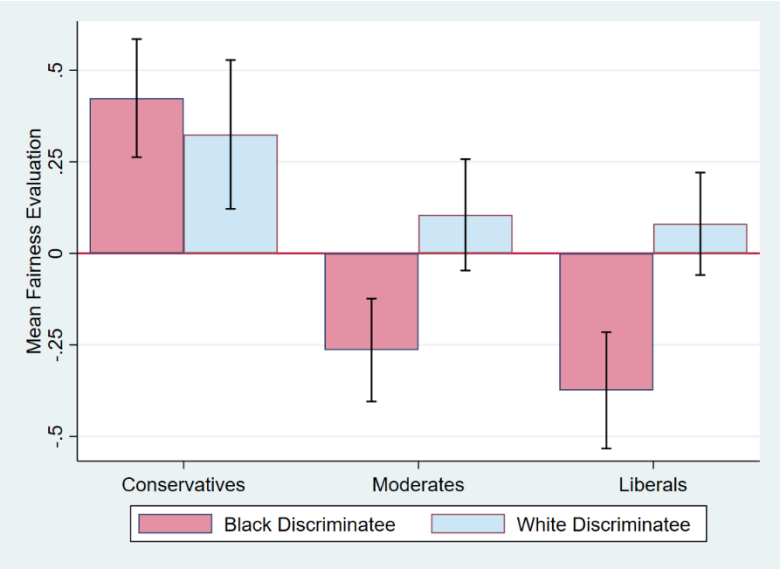
Within White respondents: *p*=.000

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

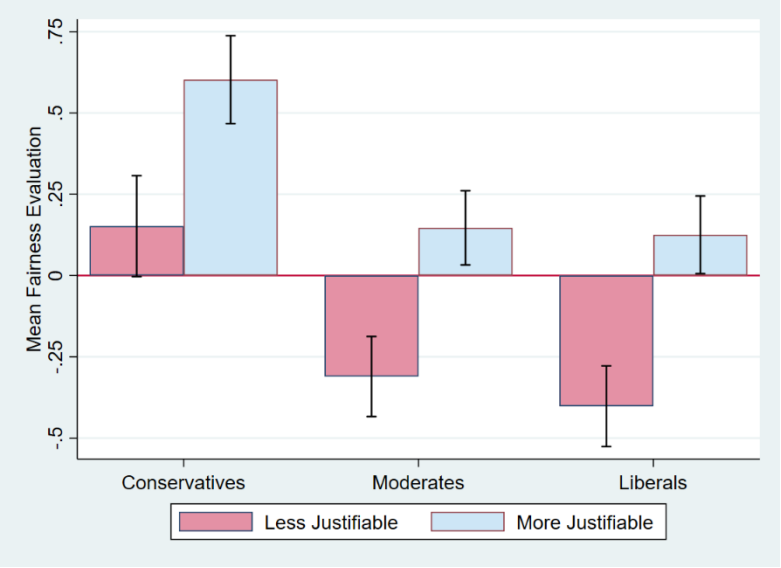
Within Conservatives: *p*=.448

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

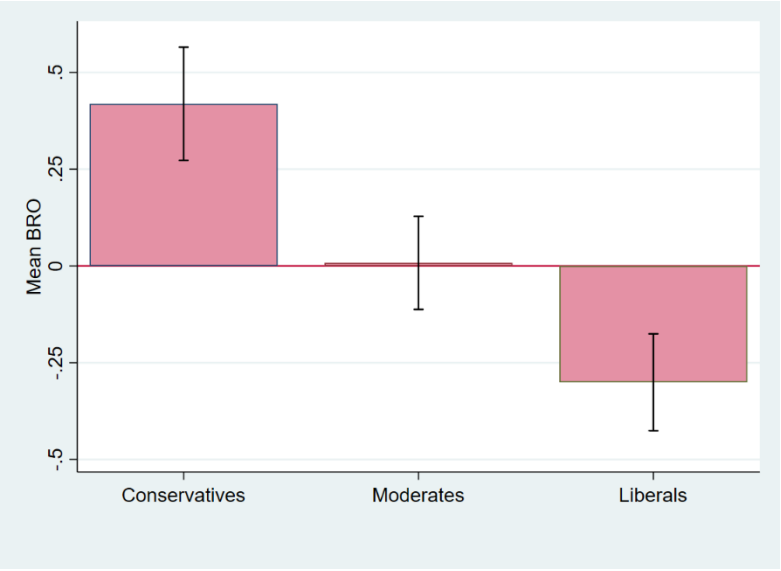
Within Conservatives: *p*=.000

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:

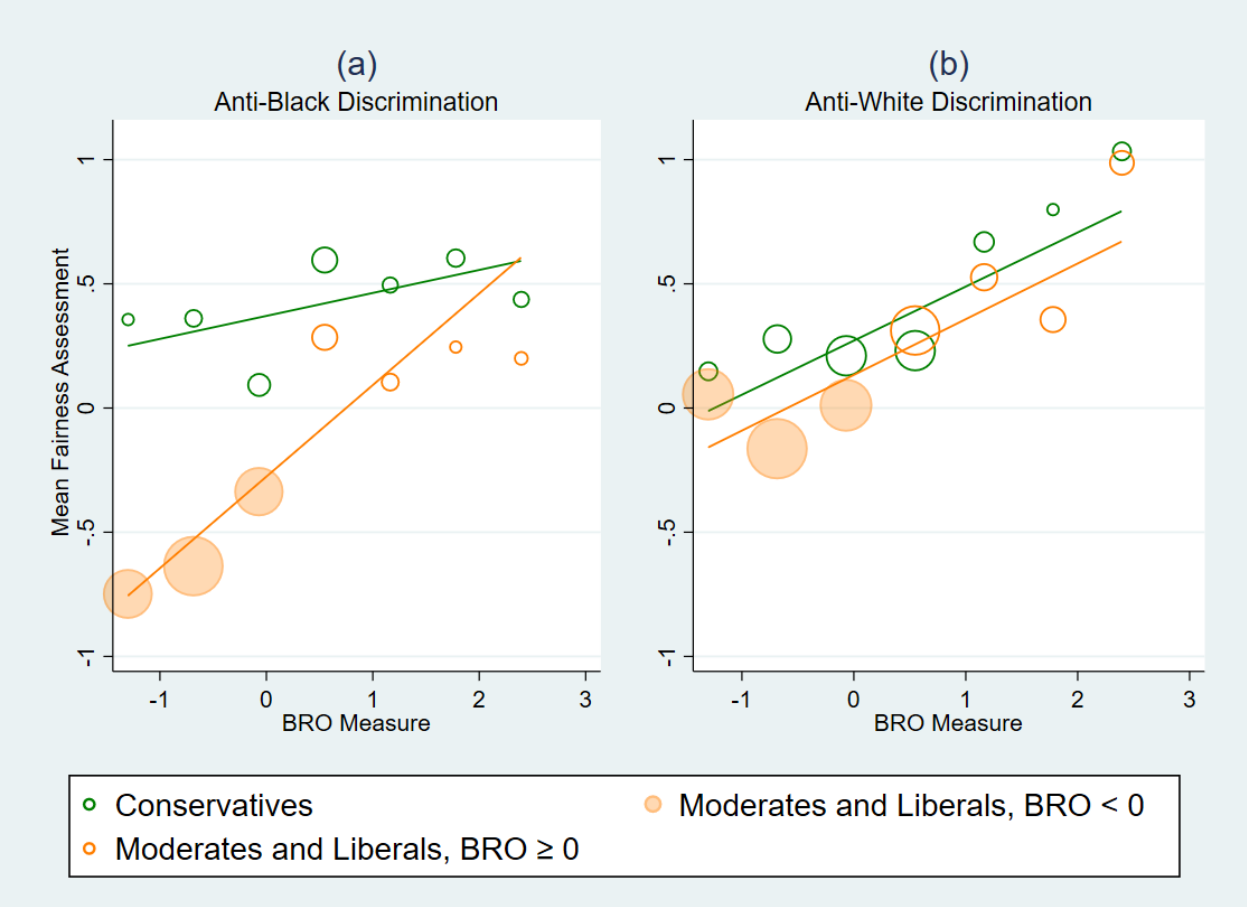
Conservatives vs. Moderates: *p*=.000

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

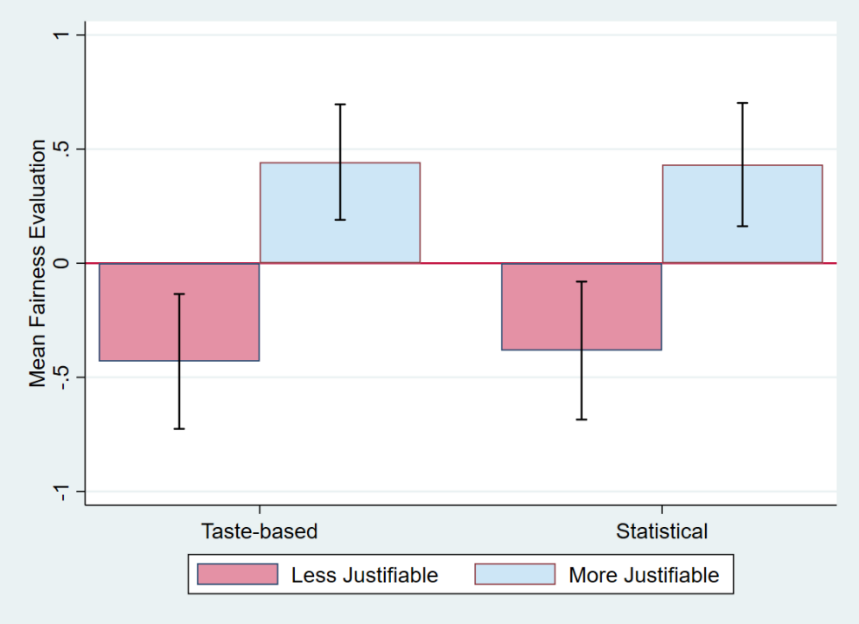
Appendix 12: Replicating the Main Figures for 'Thoughtful' Subjects Only

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)



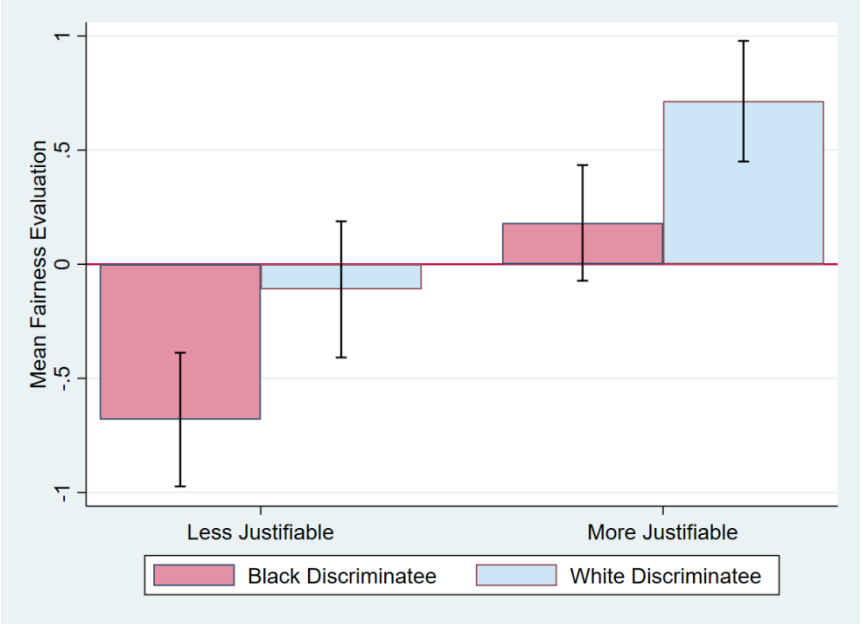
p-values:

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:

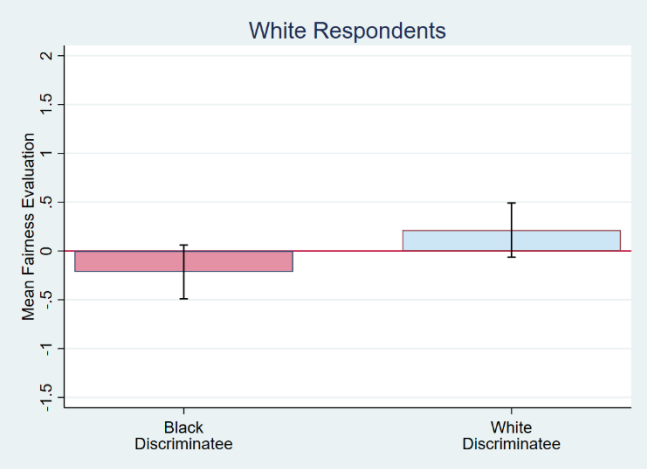
Black versus White Treatment:
Overall: *p*=.002
Within Less-Justifiable: *p*=.007
Within More-Justifiable: *p*=.004

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

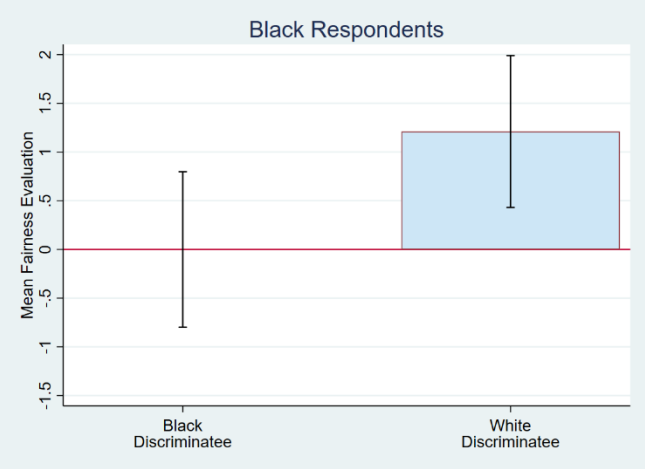
Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.861 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.825 units less fair. A test for equality of the Less versus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

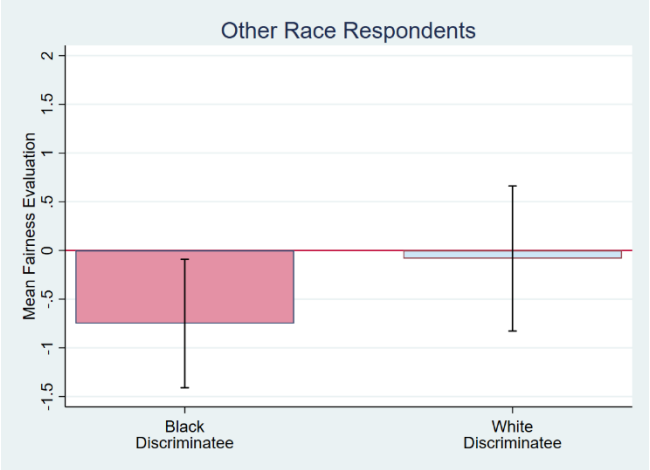
(a)



(b)



(c)

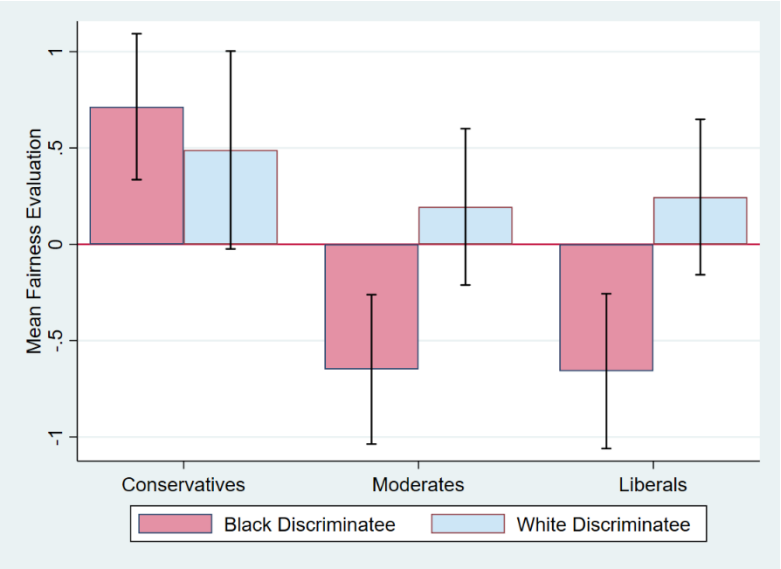


p-values:

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)

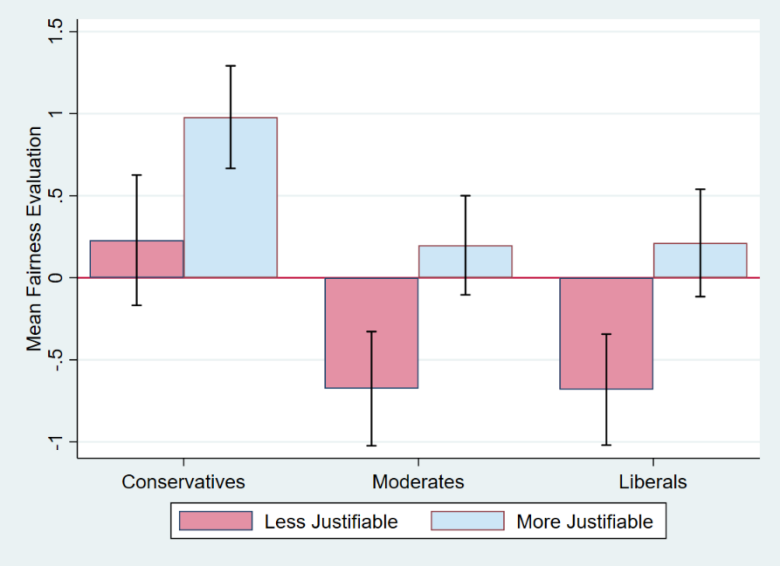


p-values:

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)

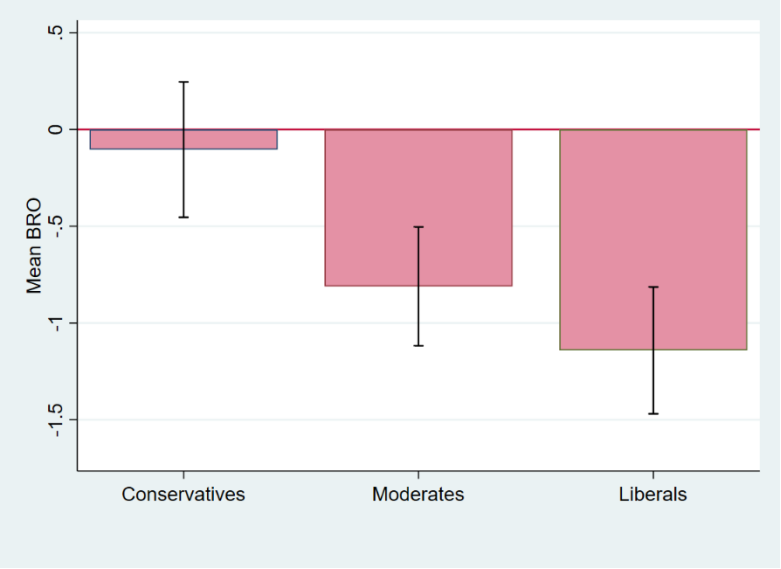


p-values:

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

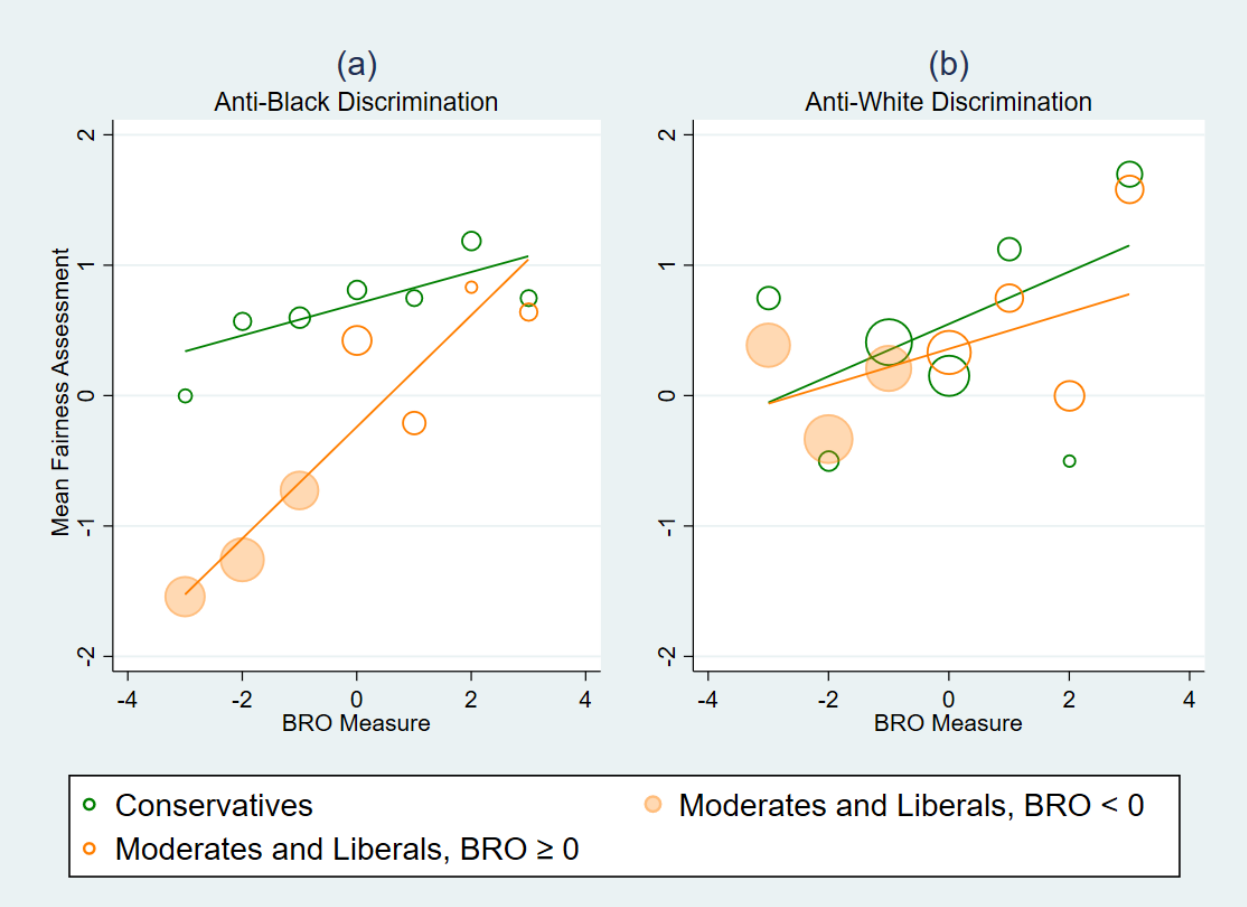
Moderates vs. Liberals: *p*=.145

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.[56] 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.[57] 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.

[56] 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.

[57] 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.) [58]

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.[59] *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.

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

[59] 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.[60]

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.

[60] 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).[61] 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.

[61] 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.