

When is Discrimination Unfair?

July 4, 2022

Peter Kuhn¹

Trevor Osaki²

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 do not distinguish between taste-based and statistical discrimination, but they react in very similar ways to other aspects of the act, 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 simple models of fairness –utilitarianism, race-blind rules (RBRs), racial in-group bias, and belief-based utilitarianism (BBU)-- and show that the latter two are inconsistent with major patterns in our data. Instead, we argue that a two-group model in which conservatives care only about race-blind rules (RBRs), while moderates and liberals care about both RBRs and utilitarian ethics can account for the main patterns we see.

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 and participants in the 2022 Trans-Pacific Labor Seminar and a seminar at Drexel university 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. 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 and political behavior (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 very little attention among economists.

To answer these questions, 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 ability of four different models of perceived fairness to account for the patterns we observe.

Our main findings are as follows. First, despite many economists' interests in distinguishing between taste-based and statistical discrimination, our subjects do not perceive a meaningful distinction between the fairness of these two types of discrimination: When faced with the same employer action (i.e., the employer rejected an equally qualified applicant because of that person's race), subjects do not care whether the decision was made because of racial animus or an expected productivity difference. Second, subjects *do* care about other aspects of employers' motivations. Specifically, taste-based 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, and 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 widely shared: they are the same across all political groups, and they do not depend on the race of the discriminatee.

Third, respondents' assessments of the fairness of discriminatory actions differ dramatically by their political orientation. On average, conservatives rate the discriminatory actions depicted in our scenarios as slightly more fair than unfair, regardless of the race of the discriminatee. Moderates and liberals rate the discriminatory acts we describe as unfair; also, in contrast to conservatives, moderate and liberals exhibit a *discriminatee race effect*: they disapprove more of anti-Black than anti-White discrimination.

Fourth, among the four models of perceived fairness we consider –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. Specifically, racial in-group bias is inconsistent with the fact that, on average, respondents from all racial groups view anti-Black discrimination as less fair than anti-White discrimination. This effect is, in fact, especially strong and significant among White respondents, though we cannot reject that its magnitude is equal across all racial groups. We reject belief-based utilitarianism –a model in which, for example, conservatives' beliefs that Black people have equal or greater economic opportunities than White people can explain conservatives' assessment of anti-Black discrimination as fair-- in part because these beliefs (which are common) cannot account for the large political gap in fairness perceptions. The predictions of the BBU model are also at odds with how respondents of all political leanings rate the fairness of anti-White discrimination.

Turning to our two remaining fairness criteria –(simple) utilitarianism and race-blind rules (RBRs)-- we argue that a two-group model in which conservatives care only about RBRs, while moderates and liberals care about both RBRs and utilitarian ethics can account for most of the patterns we see. Finally, we use within-subject treatment variation in the race of the discriminatee to assess the relative weight moderates and liberals place on the two fairness criteria they appear to care about, finding that they place roughly equal weight on utilitarianism versus RBRs when forced to choose between them.

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). Some of these authors have argued, for example, that 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 related 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 income differentials; these 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 both race-based policies like affirmative action; these beliefs also account for much of the partisan

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.

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.

divide in policy support. Informational treatments designed to change people's beliefs, however, have limited effects on policy support.

Our paper differs from these 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 various public policies. Second, and closely related, we consider a broader set of implicit fairness models that people might use to assess either actions or policies. Specifically, we will 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 of rules when mapping actions into fairness levels-- play important roles in non-consequentialist ethics such as those studied by Andreoni et al. (2019). A person's intent also plays a prominent role in criminal and civil law. In the 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 are perceived as fair or unfair.

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.

The rest of the paper proceeds as follows. Section 2 discusses the survey design, data collection, and sample characteristics. Section 3 presents some facts about fairness perceptions: How do perceptions vary with respondent characteristics, survey treatments, and the respondents' decision environment (i.e., the number and type of previous treatments encountered)? Section 4 describes four simple models of fairness and compares their implications to subjects' aggregate response patterns in Stage 1 of our survey.³ Section 5 uses within-subject treatment variation between Stage 1 and 2 of the

³ In Section 2.4 we show that question order effects matter for our *race* treatments, but not for our other two treatments (*motivation* and *justifiability*). Because the *race* treatments do not change within Stage 1 of the experiment, Sections 3 and 4 of the paper use only Stage 1 experimental data, in order to paint a clean picture of how subjects respond to all our treatments.

survey to study how subjects trade off their preferences for utilitarianism and race-blind rules when those objectives conflict. Section 6 concludes.

2. Survey Design

2.1 Overview

Before starting our survey, all our subjects are 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).⁴

Next, our respondents read and react to four vignettes that describe discrimination in a hiring decision. These vignettes randomly vary the race of the discriminator and discriminatee, as well as the *motivation* for the discriminatory action. The motivation could either reflect *taste-based* or *statistical* discrimination. Within each of these two broad categories are subcases in which discrimination is *more* versus *less* justifiable, where *justifiability* refers to how we hypothesized respondents would react. Specifically, *less*-justifiable taste-based discrimination is based on the employer’s own distaste for people of a particular race, i.e. employer discrimination. More-justifiable taste-based discrimination by the employer occurs when the employer is accommodating their customers’ distastes toward a particular race, i.e. customer discrimination. *Less*-justifiable statistical discrimination is based on low quality information – *hearsay* – about the relative productivity of workers from different racial groups.⁵ Finally, statistical discrimination is categorized as *more* justifiable if it is based on high-quality, unbiased information about the relative productivity of different racial groups for this job.

⁴ Complete instructions are provided in Appendix 1.

⁵ The *less* and *more* justifiable variants of statistical discrimination correspond to what Bohren et al. (2019) describe as “inaccurate” and “accurate” discrimination, respectively. In Bohren et al.’s case, the accuracy refers to the decisionmakers’ prior beliefs about a group’s productivity level. For example, inaccurate beliefs could be based on stereotypes.

Each vignette describes an employer, randomly named “Michael” or “Andrew,” who is making a hiring decision between a White and a Black applicant.⁶ In all cases, the (unnamed) applicants are identical in all respects except their race. As mentioned, the employer’s race can vary between vignettes, and he is always depicted as selecting the worker from his racial in-group. After each vignette, respondents are prompted to rate the fairness of the employer’s hiring decision on a seven-point scale. Since we are interested in measuring respondents’ first-order beliefs, none of these elicitation questions are incentivized.

The four scenarios encountered by each respondent are presented in two Stages. Each Stage is assigned a *race* and *motivation* treatment. The race treatments, Black and White, indicate the *race* of the discriminatee while the *motivation* treatments, Taste and Statistical, indicate the type of discrimination. Although respondents switch race and/or *motivation* treatments between Stages, respondents encounter both the *less-* and *more* justifiable subcase of either type of discrimination within each one. These *justifiability* treatments are administered in random order. Furthermore, the name of the employer also switches between the Stages.⁷ Our survey concludes with Stage 3, which asks all subjects the same background questions.

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:

⁶ Michael and Andrew appear to be among the most common male names that are relatively race-neutral. Between 2011 and 2016, they ranked in the top 2-6 names for White people and the top 6-12 names for Black people in New York City birth names. We use Michael in all our exhibits of survey instruments in the paper.

⁷ Specifically, the name of the employer is randomized (50/50) to be either “Michael” or “Andrew” in both vignettes of Stage 1. In Stage 2, the name of the employer switches to the other, unused name for all respondents.

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”.

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 changed. White scenarios were identical to Black scenarios except that the races of the discriminator (Michael) and the discriminatee are reversed. *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). To anticipate, we will find that the

circumstances summarized by these *justifiability* treatments have large effects on our subjects' fairness ratings, within both the Taste and Statistical types of discrimination.

The fact that these precise circumstances matter for fairness raises an important issue about how we can compare the perceived fairness of Taste versus Statistical forms of discrimination; essentially our comparisons hinge on the assumption that the average 'severity' or justifiability of our two Taste scenarios is the same as our two Statistical scenarios. While we cannot think of any non-tautological way to test this assumption, we designed the more-justifiable versions of our Taste and Statistical discrimination scenarios to be the mildest clear examples of each type we could think of; we also made the difference between their more and less-justifiable variants equal in magnitude from our own perspective. Our intuition on the latter issue seems to be confirmed by the data, which show a very similar justifiability gap within the Taste versus Statistical treatments.

In addition to raising sample size – while preserving the option to use only each subject's first treatment for pure cross-subject comparisons – our motivation for exposing subjects to four scenarios has two additional benefits: First, it allows us to assess the relevance of race-blind ethics in a context where individual subjects are exposed to discriminatees of different races. Second – at least to the extent that the (randomly assigned) first *race* treatment subjects encounter affects their beliefs about the experimenters' political preferences — it allows us to test for experimenter demand effects.

2.3 Common Questions, Randomization, and Representativeness

After assessing the four scenarios in Stages 1 and 2, the subjects entered Stage 3 of the survey, where they answered a common set of questions. This included a question where subjects were asked to assess the relative “economic opportunities available to Black and White people” in the United States on a seven-point scale. We also collected information on the subjects' age, education, race, gender, and political affiliation.

Turning to our randomization approach, in Stage 1 of the fairness assessments, subjects were assigned with equal probability to one of the four possible treatment combinations: SW, TW, SB, and TB (where S, T, W and B are for statistical, taste, White and Black). In Stage 2, respondents were randomly assigned to one of the three combinations they did not encounter in Stage 1. Thus, as illustrated in

Appendix A2.1, two thirds of the respondents encountered a switch in the discriminatee's race, and two thirds encountered a switch between tastes-based and statistical discrimination. Within each of the two Stages, respondents encountered the *less and more* justifiable forms of discrimination in random order.

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, 2020.⁸ 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 \$5.⁹ 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. Fifth, respondents were exposed to each vignette for at least 30 seconds before being allowed to submit their fairness assessment. Sixth, Appendices 7 and 8 re-weight all our main estimates to reflect the demographic mix in the American Community Survey and the political mix in the General Social Survey, with very similar results. Finally, Appendix P3.2 replicates some of our main results for a subset of 'thoughtful' respondents –those who took more than the median amount of time to complete the survey, again with very similar results. Some additional data cleaning resulted in a final count of 642 responses for the survey sample.¹⁰

Finally, we were concerned that the representativeness of our results would be affected by the civil unrest following the murder of George Floyd on May 25, 2020, which 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 an unusual way. To check for this possibility Figure

⁸ We re-weight our main results to match American Community Survey and General Social Survey data in Appendices 7 and 8. For additional discussions of the representativeness of MTurk samples, see Kuziemko et al. (2015) and Arechar et al. (2017).

⁹ 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.

¹⁰ Additional details on these procedures, and summary statistics on the sample's characteristics (race, gender, education, political orientation, and location in the U.S.) can be found in Appendix 2.2.

A2.2.1 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 be a significant issue for our respondents.

2.4 Question Order Effects

In Appendix 3, we demonstrate 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.¹¹ Second, the order in which respondents encounter the Tastes versus Statistical and the *more* versus *less justified* scenarios does not affect their fairness ratings. 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 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.¹²

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 were considerably more tolerant of anti-Black discrimination in Stage 2 if they were randomly exposed to a White discriminatee in Stage 1 (compared to a Black discriminatee in Stage 1). In the next two Sections of the paper (3 and 4), we will eliminate the influence of 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

¹¹ 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.

¹² 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.

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.¹³ To account for within-subject correlation of responses, all standard errors are clustered by subject.

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

Appendix 2.3 documents how the mean perceived fairness of discriminatory acts varies with respondents’ characteristics. To maximize our sample size for these comparisons, we pool responses across both scenarios each subject encountered in Stage 1, regardless of the treatment that was assigned (*motivation, justifiability, or race*). In short, Appendix 2.3 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; Appendix 2.4 explores this correlation and argues that higher levels of education mostly reflect a higher ‘set point’ for all fairness assessments, with more-educated

¹³ 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 the four scenarios each person encountered, although the standard errors are somewhat higher. See for example Figure A6.1.1, which replicates Figure 1 using first-scenario data only.

individuals rating *all* the scenarios they encounter as more fair than less-educated individuals, regardless of the race of the discriminatee and other respondent characteristics (such as political orientation).¹⁴

Finally, Appendix 2.3 shows that respondents' political leaning has sizable effects on 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). Mean fairness assessments across U.S. political *party* preferences (e.g., Democrats versus Republicans) exhibit similar patterns, though there is a statistically insignificant non-monotonicity, with Independents being more opposed to discrimination than Democrats. Since the Independent group could include people with both extreme right- and left-wing orientations, we use conservative-liberal leaning rather than party affiliation to categorize respondents' political preferences in the remainder of the paper.¹⁵

3.2 Treatment Effects on Fairness Assessments

In Figure 1, 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.¹⁶ 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.”¹⁷ 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 1, 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

¹⁴ Specifically, Appendix 2.4 shows 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. Also, more-educated respondents' greater tolerance of the discriminatory acts we depict is not confined to discrimination against a particular race: in fact, the phenomenon applies equally to anti-Black and anti-White discrimination.

¹⁵ All the results by political party are very similar, with occasional non-monotonicities similar to Figure A2.3.1(e), where Independents appear to be to the left of Democrats.

¹⁶ Figure 2 shows that the effects of *justifiability* are virtually identical for White versus Black discriminatees.

¹⁷ As noted, the subjects saw these verbal descriptions, associated with the numerals 1 through 7.

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.¹⁸ In contrast, the less 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 2 we turn our attention to the *race* treatment—i.e. the race of the person who was discriminated against. Motivated by Figure 1 (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 2 shows that respondents view the same discriminatory acts more negatively when they are directed at Black than at White job applicants. These differences are substantial in magnitude, amounting to about 0.5 fairness units or 0.263 standard deviations, and are highly statistically significant ($p = .002$ and $.000$ within the *less* versus *more* justifiable forms of discrimination, respectively).

3.3 Discriminatee Race Effects by Respondent Race and Political Orientation

While the effects of the *race* treatment shown in Figure 2 are interesting, these effects may not be the same for all types of respondents. For example, one might expect Black respondents to react more negatively than White respondents to discrimination against Black job applicants. To explore this issue, Figure 3 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 do however suggest that both groups respond to the race of the discriminatee in much the same way as White respondents

¹⁸ 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.

do.¹⁹ In sum, Figure 3 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 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 4 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.²⁰ A final striking finding from Figure 4 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 take advantage of this similarity to simplify our analysis by comparing just two political groups—conservatives versus moderates/liberals.

4. Understanding the Facts-- 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 ultimately reject two of the models –racial in-group bias and BBU—as relevant to our context because their predictions are starkly inconsistent with some key facts. We conclude by estimating a regression model that quantifies the relative importance of the two remaining models –utilitarianism and

¹⁹ Interestingly, Figure 3 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 somewhat imprecise.

²⁰ Both confidence intervals are bounded above zero (“neither fair nor unfair”).

RBRs-- for each of our two political groups. 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.

4.1 (Simple) Utilitarianism

Utilitarian models of fairness share two main components, the first of which is that fairness depends on outcomes, not on intentions or justifications. In our case, this means that fairness depends on the consequences of the employer's choice – that one applicant got the job offer and the other did not— and not on the reasons why the employer made that choice. Second, utilitarian welfare criteria assign higher levels of fairness to outcomes that redistribute wealth or opportunities from people with higher incomes to those with lower incomes. In the context of our survey, utilitarian respondents should therefore assign lower levels of fairness to acts of discrimination against Black people, whose incomes are on average significantly lower than White peoples'. Utilitarian respondents should not distinguish between Taste-based versus Statistical discrimination, nor among the less or more justifiable forms of each.

We refer to the type of utilitarianism described in this Section as 'simple' because, In addition to reducing income disparities between groups, utilitarians might also wish to equalize other features of the economic environment, including the economic *opportunities* available to racial groups. As we shall document, however, our respondents' perceptions about racial differences in opportunities vary dramatically. (See also Davidai and Walker (2021), Kraus et al. (2017), and Kraus et al (2019) who document substantial misperceptions of racial opportunity gaps.) For this reason, we frame our 'simple' utilitarian model in terms of racial *income* differences, because Black Americans' incomes are indisputably lower. We will address the effects of perceived opportunity gaps on fairness assessments under the heading of *belief-based utilitarianism*. In Section 4.4.

Turning to the evidence on simple utilitarianism, as already shown in Figure 2 respondents in general *do* view discrimination against Black applicants more harshly than discrimination against White applicants. That said, Figure 4 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 (simple) utilitarian preferences may play a role in moderates' and liberals' fairness assessments, but

are not consistent with conservatives' fairness statements. Utilitarianism also cannot account for the large justifiability effects that are documented in Figures 1 and 2.

4.2 Racial in-group bias

Like utilitarianism, most models of in-group bias are *consequentialist* in nature: they assign fairness to actions based on the action's consequences, not on the actor's intentions.²¹ The difference is that –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. Recent evidence of in-group bias includes work by Luttmer (2001), Chen and Li (2009), and Fong and Luttmer (2009, 2011). In our experiment, which studies a survey respondent's assessment of discriminatory acts against a (fictitious) employee, respondents who exhibit racial in-group bias should view those acts as less fair when the fictitious discriminatee shares the respondent's race.²² As Figure 3 has already shown, we do not have the statistical power to test these predictions for the respondents in our Black or Other racial categories.²³ 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.²⁴ 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

²¹ Chen and Li (2009) review a variety of models that could account for in-group bias in allocation decisions.

²² 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.

²³ 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.

²⁴ Figure A6.2 replicates Figure 4 (which shows discriminatee race effects by political orientation) for the subset of our respondents who are White. For White conservatives, Figure A6.2 shows weak evidence that is consistent with racial in-group bias: They rate discrimination against Black people as 0.405 units *more* fair than discrimination against Black people, but this difference is not statistically significant at conventional levels ($p = .134$).

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.²⁵ Finally, rules-based models of fairness are *race-blind* when the rules that assign fairness to actions 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 take a 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. A specific type of discriminatory act should be seen as equally fair or unfair, irrespective of the job applicant’s race.

The fairness ratings of our respondents are consistent with the use of race-blind rules (RBRs) in at least three ways.²⁶ First, the effects of our *justifiability* treatments in Figure 1 strongly support the idea that respondents care about the employer’s motivation for discriminating against a job applicant. Second and more strikingly, Figure 2 also 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 ($p=.679$). 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.²⁷

²⁵ 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.

²⁶ In Section 5.2, 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.

²⁷ Within Black Discriminatees, Taste-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 $p = .319$.

In all these cases a change in the motivation for an action has the same effect on the action's perceived fairness, regardless of the race of the discriminatee.

A reasonable concern with these striking justifiability results is that they might be driven by the fact that our subjects always encounter both the *less* and *more* justifiable forms, one after the other, within each Stage. We also draw subjects' attention to the differences between the two successive scenarios in the survey instructions. Thus our respondents might have paid close attention to the *relative* fairness ratings they assign to less- versus more-justifiable scenarios. As discussed in Section 6's robustness analysis, we explored this issue by replicating Figures 1 and 2 using only the first scenario each respondent encountered. These estimates use cross-sectional variation only to identify the effects of *justifiability*, and are uncontaminated by any possible desire to be consistent with previous fairness assessments the respondent made. Remarkably, the results are essentially indistinguishable from Table 1. We conclude that subjects' desires to maintain a consistent ranking of these forms of discrimination (perhaps because of experimenter demand effects) are not responsible for these patterns.

Reconciling our justifiability results with a purely consequentialist model of fairness is challenging, in part because the consequences of the employer's discriminatory act – being denied a job offer — are held constant in Figure 1.²⁸ Still, a consequentialist might argue that the less-justifiable forms of discrimination inflict a greater amount of *psychological* harm on the discriminatee. For example, it might be more painful to be rejected because the boss didn't like your race than because his customers didn't like your race. If so, a utilitarian welfare criterion *could* justify a larger fairness penalty for a poorly justified act of discrimination. That same utilitarian criterion, however, should attach a larger fairness penalty to poorly justified discriminatory acts against Black discriminatees, compared to Black discriminatees (because the acts inflict psychological harm on a lower-income group). As Figure 2 showed, this is not the case: The low- versus high-*justifiability* gaps for White versus Black discriminatees are very similar (-0.953 versus -0.898 fairness points respectively with $p = .679$ for a test

²⁸ While the material consequences of not being hired could vary with the discriminatee's race (because of differences in outside labor market options), Figure 1 varies only the *reasons* for not being hired: discriminatee race is balanced between the motivation and justifiability treatments due to random assignment. Figure 2 also shows that the justifiability treatments have similar effects even after we condition on discriminatee race.

of equality). Thus, even introducing arbitrary *psychological* consequences that vary with the intent of the act cannot rescue a consequentialist explanation of the *justifiability* treatment effects we observe.

We conclude our discussion of race-blind rules by noting that the preceding evidence in their favor applies on both sides of the U.S. political divide. We show this explicitly in Figure 5, 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 does 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 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, despite the fact that 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).²⁹ 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.³⁰ From a modeling perspective, this 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

²⁹ 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.

³⁰ The data in our survey do 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.

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 appears to support BBU is presented in Figure 6, which draws on our survey's measure of Black people's relative economic opportunities (BRO-- from the common questions at the end of the 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" to "much more opportunity". Figure 6 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 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.³¹

To assess whether BRO differences can actually account for the partisan gap in fairness assessments, panel (a) of Figure 7 shows respondents' fairness ratings for discrimination against Black applicants by BRO categories, separately for conservatives and moderates/liberals. If BRO accounts for the large partisan gap, we should see little or no partisan gap *within* the BRO categories; the partisan gap should be explained, instead, 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 7 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

³¹ 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 contribution relative to these papers is that we study the fairness of individual (discriminatory) actions (not public policies) and we demonstrate the key role of the intentions behind an action in determining its perceived fairness.

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).³²

A third and even more surprising piece of evidence against the “BRO hypothesis” emerges from panel (b) of Figure 7, 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) quite predictive of respondents’ assessments of the fairness of discriminatory acts, 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. Instead, it appears that high levels of BRO may be better understood as a marker of a political attitude that is accepting of discriminatory acts, *regardless of the recipient of the act*. While this attitude is more common among conservatives than other political groups, Figure 7 shows that it is held by some members of all political groups.³³

³² This partisan gap at the bottom of the BRO distribution is highly statistically significant. Both within subjects who have BRO levels of -3, and within subjects who have BRO levels of -2, the partisan gap is significant at $p=.000$.

³³ This result complements Haaland and Roth (2021), who randomly administer an information treatment that successfully reduces the partisan disparity in beliefs about Black people’s relative economic opportunities. This induced change in beliefs, however, does not reduce the partisan gap in their subjects’ support for pro-Black economic policies.

4.5 Quantifying the Relative Roles of Utilitarianism and RBR preferences

Motivated by the lack of support for racial in-group bias and BBU in our data, in this Section we estimate a simple regression model of respondents' fairness assessments that allows respondents to care about just two forms of fairness: utilitarianism and race-blind rules (RBR). In this model, respondents' fairness assessments, $Fair_{it}$, can be written as:

$$Fair_{it} = \alpha + \beta_1 Low_{it} + \beta_2 Black_{it} + \beta_3 (Low_{it} * Black_{it}) + \epsilon_{it} \quad (1)$$

where Low_{it} indicates exposure to a low-*justifiability* scenario and $Black_{it}$ indicates that respondent i was exposed to a Black discriminatee in the t^{th} scenario i encountered. In equation (1), β_1 (the additional fairness associated with low-*justifiability* acts of discrimination) quantifies respondents' concerns with the discriminator's *intentions* (when the discriminatee is White); β_2 (the additional fairness associated with a Black discriminatee) quantifies the strength of utilitarian preferences (in high-*justifiability* scenarios). Finally, β_3 tests the extent to which subjects' responses to the discriminator's intentions are race neutral: If $\beta_3 = 0$, respondents attach the same fairness penalty to a low-*justifiability* action, regardless of the racial identities of the people involved. Notice that equation (1) allows individual respondents to care about both race-blind rules and utilitarian concerns, and allows us to measure the relative strength of those concerns. Motivated by the robust lack of fairness differentials between Taste-based and Statistical discrimination, equation 1 does not distinguish between these forms of discrimination to simplify the presentation.

Estimates of equation (1) are provided in Table 1, separately for the sample as a whole, for conservative respondents, and for moderate plus liberal respondents. Like all the analyses in Sections 3 and 4, Table 1 uses data from Stage 1 of the survey only; thus, we have two observations per respondent ($t \in (1,2)$). Focusing first on conservatives, column 3 of Table 1 –which omits the interaction term to give us summary measures of overall effects-- shows that these respondents use some rules-based ethics: compared to more-justifiable discriminatory acts, less-justifiable acts are 0.865 units (or 0.452 standard deviations) less fair. In contrast, there is no evidence for any utilitarian concerns among conservatives: The race of the discriminatee has no detectable effect on conservatives' fairness assessments. Finally,

although the standard error is high, the small point estimates of β_3 (the interaction term) in column 4 indicate that conservatives implement their personal fairness rules in a race blind way.

Turning to moderates and liberals, column 6 of Table 1 shows that they also apply rules-based ethics: compared to more justifiable discriminatory acts, less-justifiable acts are 0.947 units (or 0.495 standard deviations) less fair. This penalty is very similar to conservatives', and the two estimates are statistically indistinguishable ($p = .600$). In stark contrast to conservatives, however, liberals exhibit strong utilitarian preferences: The same act of discrimination is 0.790 units (or 0.413 standard deviations) less fair when it is directed at a Black than a White job applicant. Notably, as defined here, these measures of the strength of moderates/liberals' utilitarian versus RBR preferences are roughly equal in magnitude.³⁴ Finally, the point estimate of β_3 in column 6 indicates that –like conservatives-- liberals and moderates implement their fairness rules (with respect to the nature of the discriminatory act) in a race blind way.³⁵ As noted earlier, the similar way in which the two political groups react to the specific circumstances of a discriminatory act (i.e. β_1 and β_3 in equation 1), appears to describe some important common ground in these groups' beliefs about what is fair in the labor market.

5. Learning from Order Effects: Experimenter Demand Effects and the Utilitarianism-RBR Tradeoff

In Section 4.5 we argued that –while conservatives' fairness assessments appear to be influenced only by race-blind rules (RBRs)—moderates and liberals seem to care about both RBRs and utilitarian criteria. In Section 2.4 we showed that subjects' fairness assessments exhibit order effects: their Stage 2 ratings depend on race of the discriminatee they encountered in Stage 1. In this Section, we exploit these order effects to study how liberals and moderates reconcile the two fairness criteria they care about – utilitarianism and RBRs—when those criteria conflict.

³⁴ We recognize that comparing discriminatory acts of different justifiability versus acts against different races is not a natural metric; the main goal of the Table 1 regressions is to quantify respondents' reactions to two different fairness determinants in a multiple regression context that allows both to operate simultaneously.

³⁵ A test of equality for the estimates of β_3 between both groups is also statistically indistinguishable ($p = .447$).

To accomplish this goal, we proceed in three steps. First, we exploit data from Stage 3 of the experiment to rule out a plausible form of experimenter demand effects as an explanation for the order effects we observe. Second, we document that these order effects are only present among moderate and liberal respondents. Finally, we use a simple model of reporting behavior, combined with random assignment of the *race* treatment to interpret moderates and liberals' order effects as a compromise between utilitarianism versus RBRs, and to estimate the relative weight moderates and liberals place on those two criteria when they conflict. The key assumption underlying our approach is that respondents only become conscious of their desire to be race-blind when they encounter discriminatees from a second racial group.

5.1 Experimenter Demand Effects

Intuitively, the order effects we observe are that subjects have a less negative view of anti-Black discrimination in Stage 2 if they encountered a White discriminatee instead of a Black discriminatee in Stage 1.³⁶ One potential explanation of such a response is an experimenter demand effect of the following form: 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.

To probe the plausibility of this demand-effects model, Appendix 4 argues 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'

³⁶ We also find that subjects' Stage 2 assessments of anti-White discrimination are more negative if they encountered a Black discriminatee in Stage 1, although this difference is not statistically significant.

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 report their own political leanings as more conservative.

In Appendix 4, 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 order effects we observe.

5.2 Race Treatment Order Effects are Absent among Conservatives

Taking it as given that our respondents' political leanings are honestly reported, we now establish one additional fact concerning the *race* treatment order effects in our data: these order effects are absent among conservatives. Specifically, Appendix 5 replicates Figure A3.4.1—which showed that respondents' Stage 2 fairness assessments depend on the *race* treatment they encountered in Stage 1—separately for conservative respondents versus moderate / liberal respondents. It shows that there are no such order effects for conservatives: Regardless of discriminatee race they encountered in Stage 1, conservatives view discrimination as a little more fair than unfair (about +0.5 on a scale from -3 to +3) in Stage 2. Moderate and liberal respondents, on the other hand, exhibit a more pronounced version of the aggregate order effects documented in Figure A3.4.1: Moderates' and liberals' Stage 2 fairness assessment of anti-Black discrimination are much milder if they encountered a White discriminatee (as compared to a Black discriminatee) in Stage 1 $p = .009$).³⁷ Motivated by this fact, we restrict our attention to moderate and liberal respondents in Section 5.3, where we propose an interpretation of the *race* treatment order effects in our data.

³⁷ The race treatment a moderate/liberal subject received in Stage 1 does not have a statistically significant effect on the subject's ratings of discrimination against White people in Stage 2.

We conclude by noting an additional implication of the fact that order effects are absent among conservatives. Specifically, this fact makes the “experimenter demand effects” hypothesis examined in Section 5.1 even less plausible. For experimenter demand effects to explain our results, 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 sum, experimenter demand effects would need to take a very special form --only affecting one of our Section 3 questions about beliefs and political attitudes, and only affecting one political group-- to account for the order effects in our data.³⁸

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

In Section 4.5, we used Stage 1 data to argue that moderate and liberal subjects may have utilitarian motives, but also appear to derive utility from race-blindness when making fairness assessments. If that is the case, subjects who are exposed to both Black and White treatments (i.e. treatment switchers) face a conflict between utilitarianism and race-blindness. 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), and respecting their utilitarian desire to 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 “true” ratings in a setting where they don’t need to consider race-blindness (B_i^* and W_i^*).

³⁸ In contrast, note that the ‘tradeoff’ model in Section 5.3 has a ‘built-in’ explanation for the absence of order effects among conservatives: Because conservatives exhibit no utilitarian concerns (as is demonstrated by their Stage 1 reports) conservatives *never face a conflict* between utilitarian preferences and a desire to be race neutral. Thus, the model itself predicts that order effects should only be present among groups (like moderates and liberals) whose Stage 1 responses indicate that they care about fairness criteria *other* than RBRs.

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 true rating and making the same report as in Stage 1 (being race-blind).³⁹

Using this notation, White-to-Black treatment switchers have the option of reporting their true rating of discrimination against the second racial group they encounter ($B_i^2 = B_i^*$), assigning the same rating they assigned (to the other race) in Stage 1 ($B_i^2 = W_i^1$), 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 ‘true’ 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-to-W switchers), their sample means \bar{B}^* and \bar{W}^* in any fixed population (such as moderates and liberals) 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,

³⁹ 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.

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

After restricting the sample to moderate and liberal respondents, Equations (3) and (4) can then be (separately) solved for α , yielding $\alpha = 0.44$ for the White-to-Black switchers and $\alpha = 0.62$ for the Black-to-White switchers.⁴⁰ Thus, W-to-B switchers' Stage 2 ratings of anti-Black discrimination are (slightly) closer to strict race-blindness ($\alpha = 0$) than to a pure utilitarian assessment ($\alpha = 1$). B-to-W switchers, on the other hand, act as if they place slightly more weight on their utilitarian 'truth' than on race-blindness. That said, we cannot reject equal weight on both objectives ($\alpha = 0.5$) for either type of switcher ($p = .678$ and $p = .423$ for W-to-B and B-to-W switchers, respectively).⁴¹

Additional hypothesis tests allow us to reject other meaningful values of α , however. Specifically, for W-to-B switchers we can reject both $\alpha = 0$ (100% weight on race-blindness) and $\alpha = 1$ (100% weight on the utilitarian 'truth') [$p=.003$ and $p=.007$ respectively]. In other words, when moderates and liberals encounter White and Black discriminatees in that order, their second fairness assessment places strictly positive weight on *both* utilitarian preferences and race-blind rules. For B-to-W switchers we can reject pure race-blindness ($p=.000$), but we cannot quite reject 100% weight on utilitarian preferences ($p=.067$). In sum, the *race* treatment order effects we observe among our moderate and liberal 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 roughly 'split the difference' between these two objectives when making their fairness assessments.

6. Robustness

One striking result of our analysis is the large magnitude, statistical significance, and stability of the *justifiability* treatments: Respondents of all political orientations penalized the less- justifiable forms

⁴⁰ Our lower estimate of α (the weight placed on the utilitarian 'truth') for W-to-B switchers than B-to-W switchers is consistent with the fact that *race* treatment order effects were only statistically significant for the former group.

⁴¹ The 95% percent confidence intervals for α are [0.217,0.706] and [0.171,0.644] for W-to-B and B-to-W, respectively.

of discrimination (relative to more-justifiable forms) by the same amount, irrespective of the discriminatee's gender. A possible concern with this result is the fact that the 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 ensure they assign lower fairness rankings to the scenario that feels less justifiable. As noted, Appendix A6.1 addresses this issue by replicating Tables 1 and 2 using only the first 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 3 of the paper showed that on average, White respondents object more to acts of anti-Black than anti-White discrimination, and argued on that basis that racial in-group bias does not play a major role in explaining our subjects' fairness assessments. Given the large partisan divide in our data, however, this raises the question of whether racial in-group bias may still be present among a subset of White respondents. To address this issue, Appendix 6.2 replicates Figure 3 for conservative respondents only. Interestingly, the discriminatee race effect *does* switch sign in this group, relative to the full sample in Figure 3: 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 ($p=0.134$).

For much of the analysis in the paper we combined moderates and liberals into a single group given their similar levels and patterns of fairness assessments. This includes Figure 7, which illustrated some strong and unexpected relationships between perceptions of Black peoples' relative economic opportunities (BRO), political orientation, and fairness assessments: For example, the acceptability of anti-White discrimination *increases* with Black relative opportunities. To see if the similarity between moderates and liberals extends to these unexpected findings, Figure A6.3 replicates Figure 7, showing separate results for moderates versus liberals. Consistent with earlier results, these two groups exhibit very similar response patterns, both of them differing substantially from conservatives' patterns.

In Section 2.3 we showed that our sample of MTurk respondents was not representative of adult Americans on a number of key dimensions. While our small sample size limits what we can do to address this issue, Appendices 7 and 8 replicate all our main analyses (Figures 1-7 and Table 1) two different ways. First, Appendix 7 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 7(a) becomes somewhat stronger and statistically significant. Similar to Figure 7, however, the slope for conservatives remains much lower than the slope for moderates / liberals.

Second, Appendix 8 replicates Figures 1-7 and Table 1 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.⁴² Despite significant differences in the political mix of the two surveys, all the main results are replicated.⁴³

An important concern in 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 issue, 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 of the paper. Briefly, Appendices P1-P3 together comprise a “populated PAP” which reports the results of the exact tests specified in the PAP. Appendix P4 then summarizes the relationship between the PAP and the paper.

Specifically, Appendix P4 begins by documenting 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 Section 4 and the main tests thereof (the models’ names have changed

⁴² 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. The ACS does not ask questions about political orientation or party preference.

⁴³ The one exception noted with the ACS weights in Appendix 7 does not occur here.

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: First to simplify the presentation, the paper mostly reports simple *t*-tests of differences in means rather than regression results. Second, we decided not to standardize our main outcome measure (fairness) because we realized that this would obscure important cardinal information. Third, motivated by the *race* treatment order effects (which we anticipated *might* be present), we restricted the sample in Sections 3 and 4 to Stage 1 responses only. Fourth, we confined Section 5.3's "learning from order effects" analysis to moderates and liberals, because these effects were strikingly absent among conservatives. Fifth, in Figure 7'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.

Finally, Appendix P4 lists the following PAP hypothesis tests that we decided *not* to include in the main paper.⁴⁴ We did not include an "actions versus identity" decomposition of fairness determinants because it seemed of limited interest; due to a lack of statistical power we did not pursue the idea of classifying subjects into types based on their responses to within-subject treatment variation; and we did not extend Appendix P3.2's replication of populated PAP results for a subset of 'thoughtful' respondents (who took more than the median time to complete the survey) to the results in the paper. This restriction had very little effect in the populated PAP so we decided the extension would add little to the paper.

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 taste-based racial discrimination. We have found, first of all, that respondents of all political leanings are indifferent to the distinction between statistical and taste-based discrimination. This may be surprising given many economists' apparent interests in

⁴⁴ The results of these tests are available in Appendices P1-P3.

distinguishing between these types of discrimination.⁴⁵ Second, we find that respondents of all political leanings *do* care about other aspects of the motivation behind a discriminatory act—specifically whether statistical discrimination is based on precise information versus hearsay, and whether the discriminator is accommodating his own tastes or those of others. Interestingly, respondents of all political leanings care about these motivational differences in very similar ways, and do so in a race-blind manner, revealing a significant domain of shared values across the political spectrum. Third, there is however a large political difference in both (a) how much respondents object to discriminatory acts in general, and (b) the extent to which their fairness ratings depend on the discriminatee’s race. Conservatives do not differentiate between anti-Black and anti-White discriminatory acts, and rate all types of discriminatory acts we depict as either neutral or slightly *more* than fair. Moderates and liberals, on the other hand, view discrimination as unfair in most cases, and object more strongly to anti-Black than anti-White discrimination.

To attempt to understand these facts, we evaluated four pre-specified models of fairness, and showed that two of them –racial in-group bias and beliefs-based utilitarianism—are inconsistent with major features of our data. We conclude, instead, that a model with two political groups and two models of fairness --utilitarianism and race-blind rules-- accounts for most of the response patterns we have identified. In this model, conservatives care *only* about race-blind rules, while moderates and liberals care about both RBRs and utilitarian ethics. When these ethics conflict (e.g. when a subject encounters a switch in the *race* treatment) moderates and liberals assign fairness ratings that place roughly equal weight on each one.

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. In this sense

⁴⁵ One motivation for economists’ interest in this question may be a perception that statistical discrimination is, indeed, less objectionable than taste-based discrimination. According to this point of view, it seems unfair to blame animus-free employers from simply using available information to hire the best worker. (See Tilcsik 2021 for quotes to this effect from well-known economists). Our results in this paper suggest that the U.S. general population may not share economists’ perception about ‘tastes’ versus ‘statistics’.

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. 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 that motivations could affect the extent to which *public policies* are seen as fair also. If this is the case, winning support for a policy involves not only crafting the details of the policy itself (who wins, who loses) but also crafting a *motivation* for the policy change that is widely perceived as fair. In addition, our results –like Alesina et al.’s (2021) and Haaland and Roth’s (2021)-- suggest 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. This group of respondents is likely to resist policies that interfere with employers’ decisions to hire and fire at will, 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 conservatives might be more receptive to policies that equalize racial opportunities. One of these are cases in which a clear rule has *not* been applied in a race-blind way; in these cases, *restoring* race-blindness should appeal to people with an RBR ethic. Second, our results suggest that conservatives, like moderates and liberals, react more negatively to race-based actions that were taken for less-justifiable reasons, like personal animus and hearsay-based evidence. Antidiscrimination policies that target these types of behaviors may thus be better received by conservatives.

Our analysis is subject to some important *caveats* and limitations. For example, all our results – including our finding that conservatives do not, on average, object to discriminatory actions-- are limited to the range of actions our scenarios depict. It seems likely that more *consequential* acts (e.g. being fired rather than not being hired, or receiving a long prison sentence) or less *justifiable* acts (e.g. ones inspired by racial hatred) would elicit stronger negative responses among all groups. Second, while we have more

than enough statistical power to test our pre-registered hypotheses, we lack the statistical power to resolve some important questions, such as whether self-identified conservatives are more accepting of anti-Black than anti-White discrimination. Constraints on sample size have also prevented us from painting a very detailed picture of the fairness assessments made by non-White Americans.

We also hasten to point out that our analysis has identified some interesting puzzles that are not resolved here. For example, while we demonstrate that race-blind rules can account for several key features of conservatives' fairness ratings, conservatives' reliance on RBRs does not explain why, on average, conservatives do not object to any of the discriminatory acts we depict. For example, instead of *never* penalizing discriminatory acts, an alternative set of race-blind rules would penalize both anti-White and anti-Black discrimination, strongly and equally. We also do not have a good understanding of why subjects of all political leanings become more accepting of (hypothetical acts of) anti-White discrimination as their perception of White peoples' relative economic opportunities becomes more pessimistic. One hypothesis in this regard is that high levels of BRO (Black people's perceived relative opportunities) are a marker for a political attitude that, for example, "too much attention is given to discrimination", and that private discriminatory acts—at least within the range depicted by our scenarios—should simply be tolerated.

Taken together, the preceding *caveats* and puzzles emphasize the fact that our analysis has only scratched the surface of the question that motivates our paper, "When is Discrimination Unfair?". Indeed, we believe that the methods in our paper could easily, and fruitfully, be applied to a wide variety of related questions. For example, discriminatory scenarios in future papers could (a) depict more *consequential* acts (such as getting a long prison sentence); (b) depict less *justifiable* motivations (such as hatred versus mild distaste); (c) vary the *context* in which the act occurs (housing markets, credit markets, judicial decisions); (d) vary the *groups* involved (such as gender, age, sexual orientation, political orientation, age, criminal and credit history); (e) change the *social environment* depicted in the scenario (for example, is the act seen by others?); (f) manipulate whether the fictitious act is *conscious* or unintended; or (g) change the respondent's *decision environment* (priming, mental bandwidth available, and whether the respondent's ratings are visible to others). Our results also suggest that economists' frequent focus on distinguishing taste-based from statistical explanations may be misplaced. Instead, it may

be more useful to focus future research more on the aspects of discriminatory acts which our results suggest are of greater concern to most people, such as the detailed motivations behind the acts and the identities of the discriminator and discriminatee.

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 Gächter, 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.
- 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.

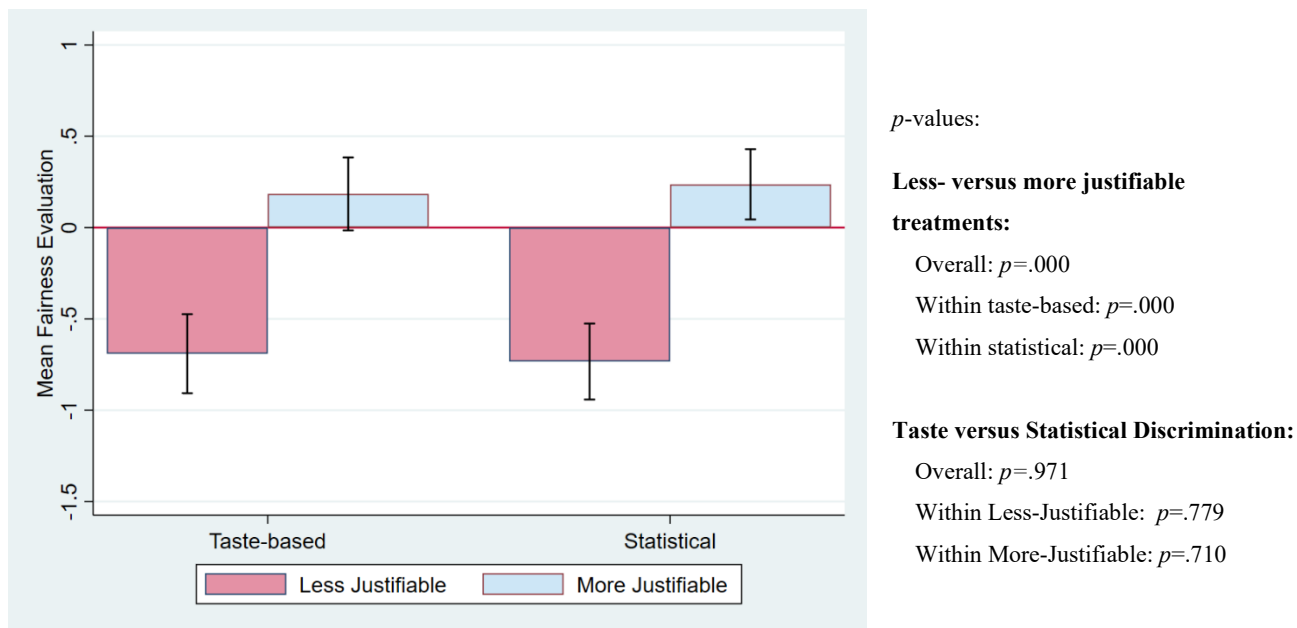
- Breza, Emily, Supreet Kaur, and Yogita Shamdasani. 2017. "[The morale effects of pay inequality.](#)" *The Quarterly Journal of Economics* 133(2): 611-663.
- Bruhlin, Adrian, Ernst Fehr, and Daniel Schunk. 2019. "[The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social 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.
- 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.
- 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.
- Sauer, Carsten. 2020 "[Gender Bias in Justice Evaluations of Earnings: Evidence From Three Survey Experiments](#)" *Frontiers in Sociology* 7(5):22.
- 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.

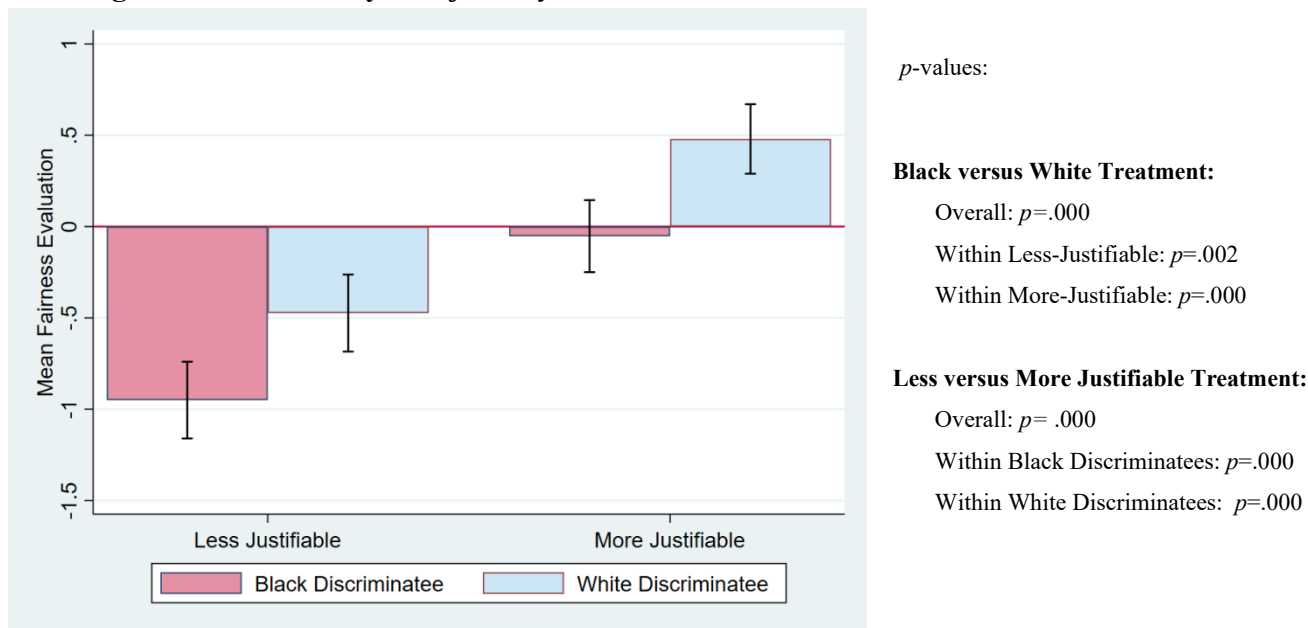
Figures and Tables

Figure 1: Fairness Ratings by Type of Discrimination and *Justifiability*



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

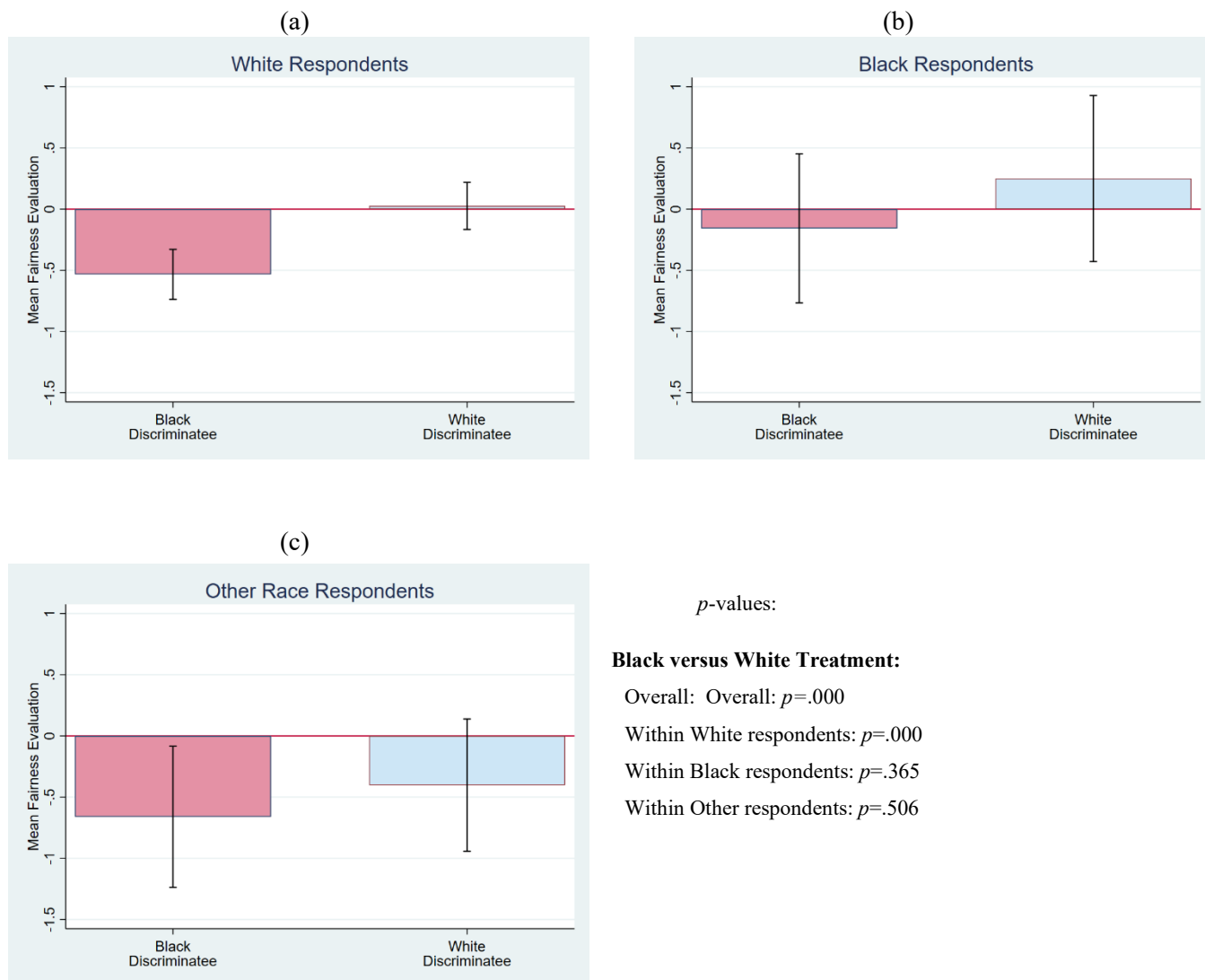
Figure 2: Fairness by *Justifiability* and Discriminatee Race



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

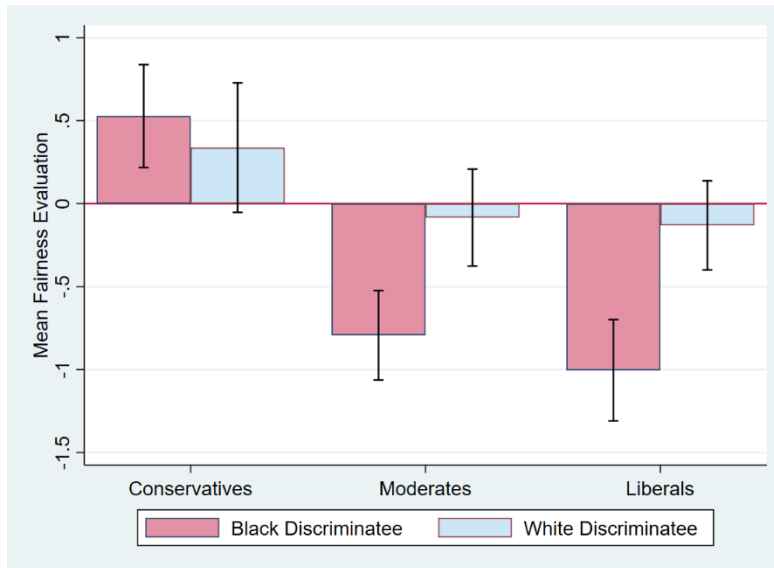
Within Black Discriminatees, less-justifiable scenarios are 0.898 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.953 units less fair. A test for equality of the Less versus More *Justifiability* Gap between the Black and White treatment yields $p = .679$.

Figure 3: Fairness Ratings by Respondent Race and Discriminatee Race



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 4: Fairness Ratings by Political Orientation and Discriminatee Race



p-values:

Black versus White Treatment:

Overall: $p = .000$

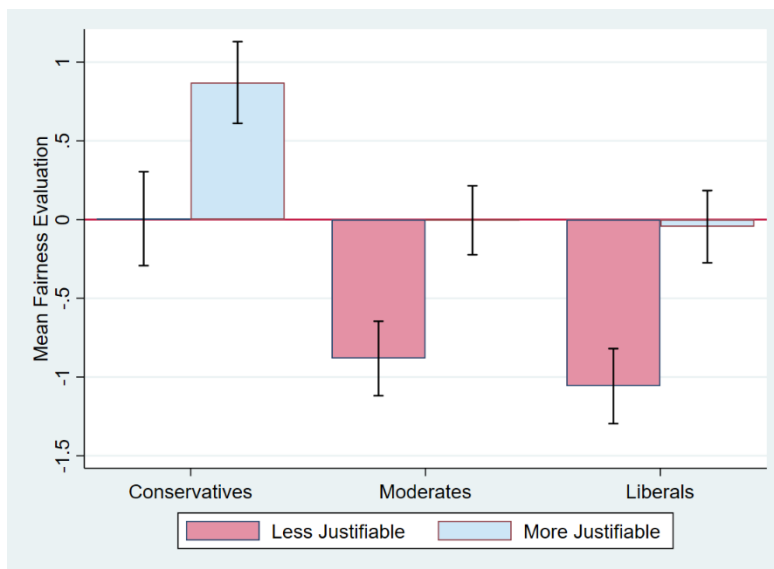
Within Conservatives: $p = .448$

Within Moderates: $p = .000$

Within Liberals: $p = .000$

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 5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent's Political Leaning



p-values:

Less versus More Justifiable Treatment:

Overall: $p = .000$

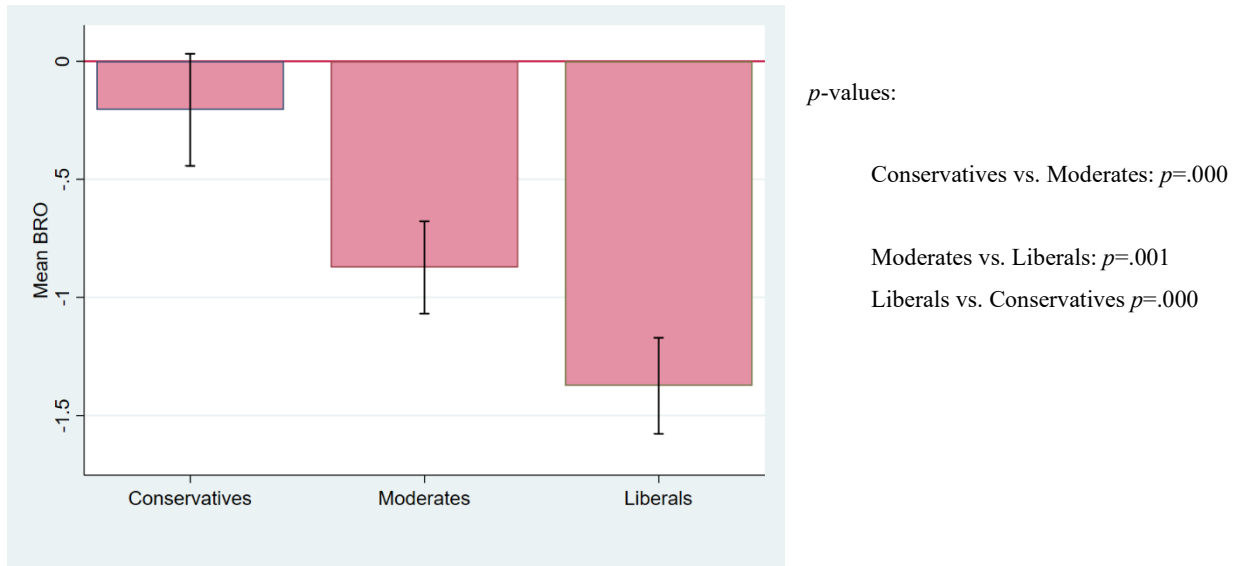
Within Conservatives: $p = .000$

Within Moderates: $p = .000$

Within Liberals: $p = .000$

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 $p = .590$.

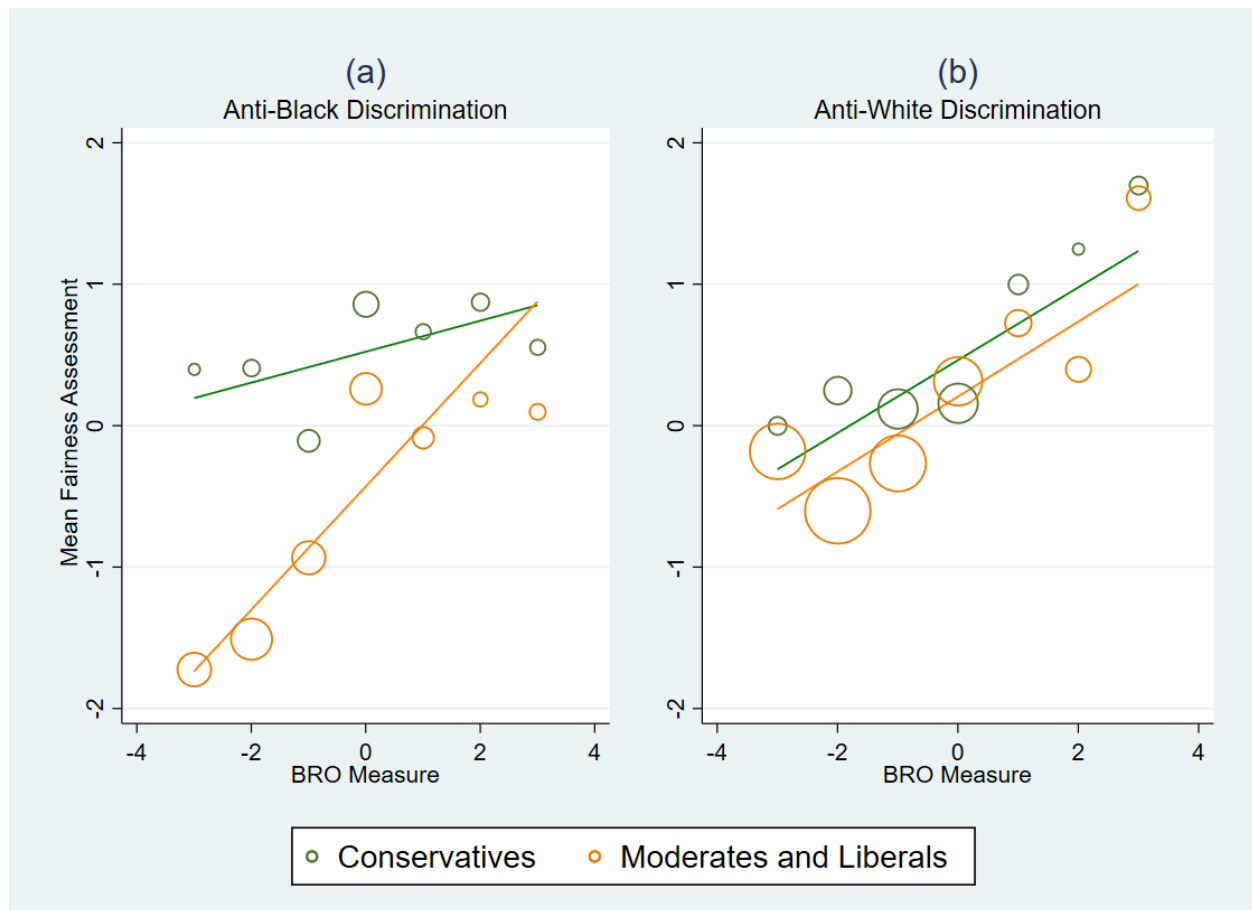
Figure 6: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning



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 $p = .577$.

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



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 = 0.109, $p = .218$
 - 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 ($p = .000$)

Table 1: Estimating the Combined Effects of Utilitarian and Rules-Based Determinants of Fairness

| | All (1) | All (2) | Conservatives (3) | Conservatives (4) | Moderates and Liberals (5) | Moderates and Liberals (6) |
|---|-----------------------|-----------------------|----------------------|----------------------|----------------------------------|----------------------------------|
| Less justifiable | -0.925*** (0.0665) | -0.953*** (0.0968) | -0.865*** (0.139) | -0.800*** (0.190) | -0.947*** (0.0757) | -1.004*** (0.112) |
| Black discriminatee | -0.505*** (0.129) | -0.532*** (0.139) | 0.190 (0.250) | 0.251 (0.266) | -0.790*** (0.143) | -0.848*** (0.157) |
| Less justifiable × Black discriminatee | | 0.0551 (0.133) | | -0.122 (0.276) | | 0.116 (0.151) |
| Constant | 0.466*** (0.0931) | 0.480*** (0.0968) | 0.770*** (0.202) | 0.738*** (0.205) | 0.365*** (0.104) | 0.393*** (0.109) |
| Observations | 1,284 | 1,284 | 340 | 340 | 944 | 944 |
| R-squared | 0.076 | 0.076 | 0.055 | 0.055 | 0.110 | 0.110 |

Notes: Regression results are based on Stage 1 fairness assessments only. 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 respondent.

Appendix (for online publication)

Appendix 1: Complete Survey Instructions and Questionnaire

This Appendix 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.

Appendix 2: Randomization, Representativeness, and Effects of Respondent Characteristics

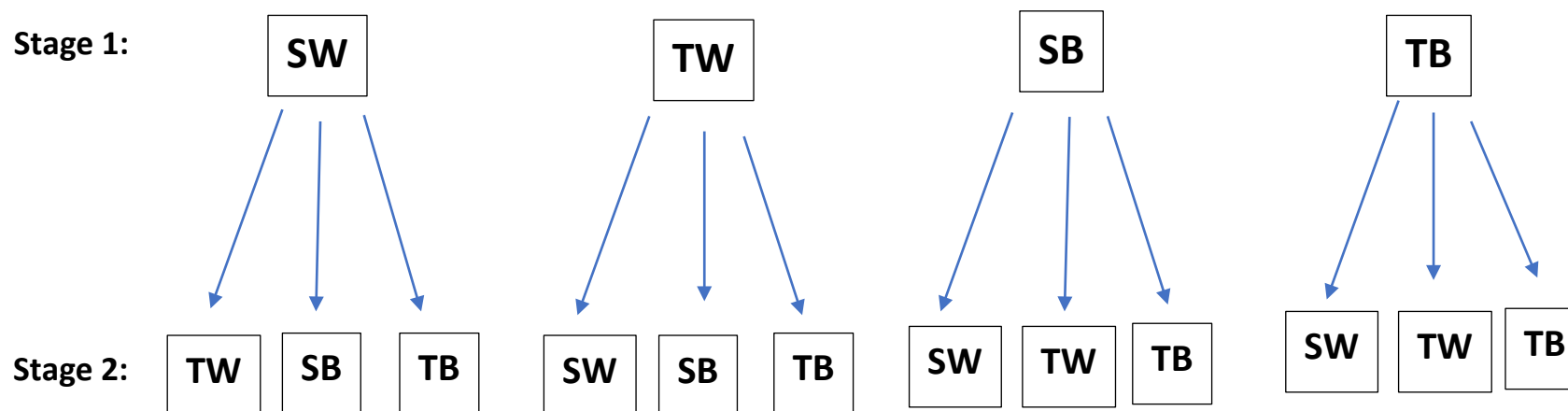
A2.1 Randomization

In Stage 1, respondents were randomly assigned with equal probability to one of these four treatment combinations: SW, TW, SB, and TB.

In Stage 2, respondents were randomly assigned to one of the three treatment combinations they did not encounter in Stage 1. (2/3 switched *race* treatments, and 2/3 switched *motivation* treatments (statistical versus tastes)

Within each Stage, respondents encountered the *less* and *more* justified forms of discrimination in random order.

Figure A2.1.1 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.

A2.2 Descriptive Statistics and Representativeness

Tables A2.2.1 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 A2.2.2 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.⁴⁶ Overall, Table A2.2.2 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 A2.2.3 and A2.2.4 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 A2.2.1 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.

⁴⁶ 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 A2.2.2 for details).

Table A2.2.1: Composition of MTurk Sample versus the American Community Survey (ACS), by Demographic Characteristics

| 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.

Table A2.2.2: 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 A2.2.3: Composition of MTurk Sample by Census Region

| CENSUS REGION | MTurk Sample (1) | 2019 ACS Sample (2) |
|--------------------|---------------------|------------------------|
| <i>Northeast</i> | 0.238 | 0.178 |
| New England | 0.028 | 0.048 |
| Middle Atlantic | 0.210 | 0.130 |
| <i>Midwest</i> | 0.189 | 0.212 |
| East North Central | 0.136 | 0.146 |
| West North Central | 0.053 | 0.066 |
| <i>South</i> | 0.394 | 0.376 |
| South Atlantic | 0.251 | 0.201 |
| East South Central | 0.047 | 0.059 |
| West South Central | 0.097 | 0.116 |
| <i>West</i> | 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.

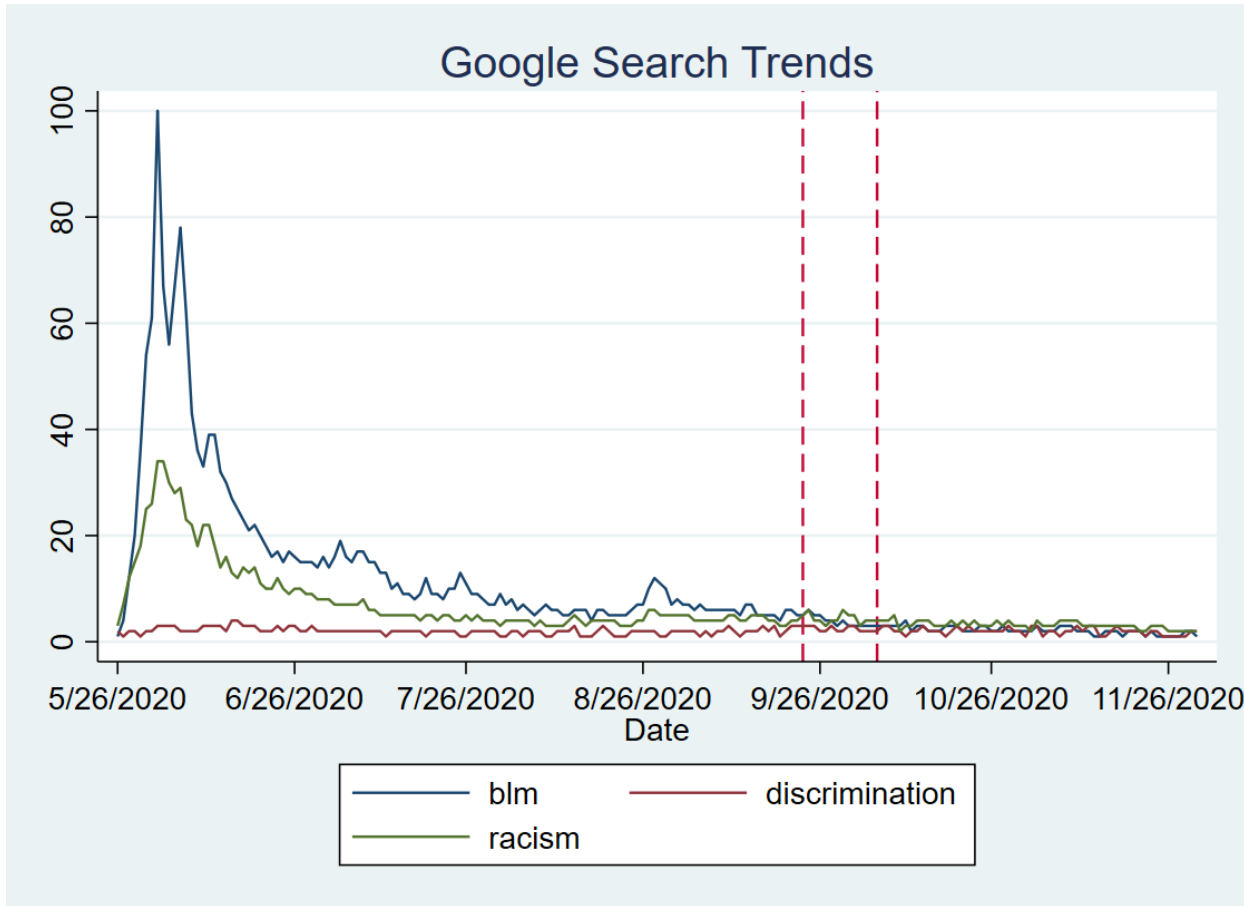
Table A2.2.4: Composition of Mturk Sample versus ACS by U.S. State (pop. exceeds 5 million)

| STATE | MTurk Sample <i>shares</i> (1) | MTurk Sample <i>Count</i> (2) | 2019 ACS Sample <i>shares</i> (3) | State Pop. <i>in thousands</i> (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 A2.2.1: Frequency of Google Searches for BLM and Related Keywords around the Survey Date

Notes: This figure illustrates trends in Google searches for keywords such as “Black Lives Matter (blm).”



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.

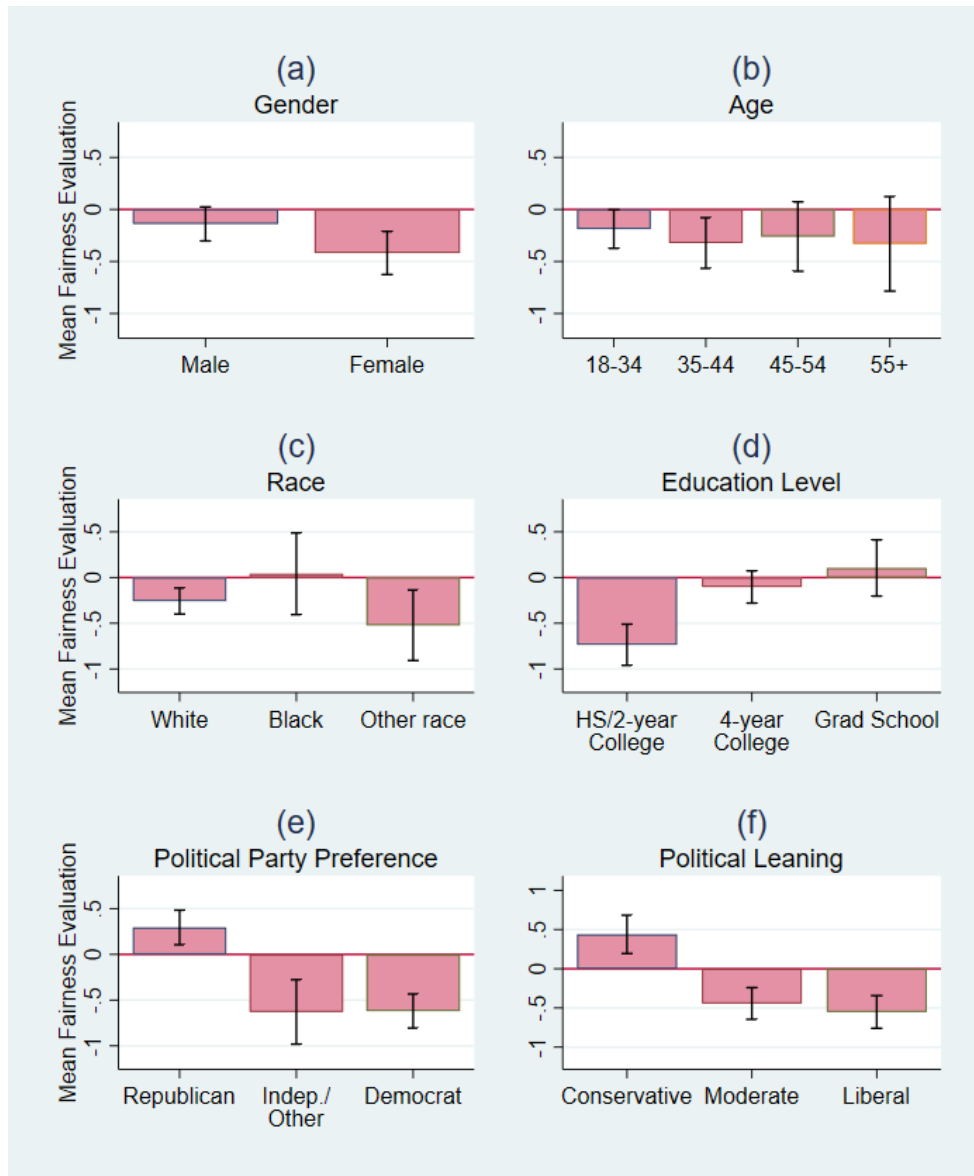
A2.3 Effects of Respondent Characteristics on Fairness Ratings

Figure A2.3.1 displays mean fairness assessments across various respondent characteristics. Overall, fairness assessments do not vary significantly with respondent age and only slightly with gender. Women view discrimination as less fair than men, but the difference in means is small ($p = .037$). Interestingly, Black respondents appear to view the same discriminatory actions as *more* fair compared to White respondents, although the corresponding difference in means is not statistically significant ($p = .204$). Furthermore, respondents from other racial groups (referred to as the “Other” group/respondents) judge discriminatory actions to be less fair compared to Black and White respondents. However, the fairness gap between the Other and White respondents is insignificant ($p = .197$). Also, contrary to our expectations, more-educated respondents view the same discriminatory actions as fairer than less-educated responses (e.g., $p = .000$ for high school plus 2-year college graduates versus 4-year college graduates).

Part (e) of the figure shows that a respondent’s political affiliation has large effects on the perceived fairness of discriminatory acts. Republicans perceive discriminatory acts to be fairer compared to Democrats and Independents (e.g., $p = .000$ for Democrats versus Republicans).⁴⁷ Similarly, part (f) shows that conservative respondents perceive discriminatory acts to be fairer compared to both independents and liberals (e.g., $p = .000$ for conservatives versus liberals).

⁴⁷ “Independents” includes a small number of respondents who claim an affiliation with a third political party.

Figure A2.3.1: Mean Fairness of Discriminatory Actions by Respondent Characteristics



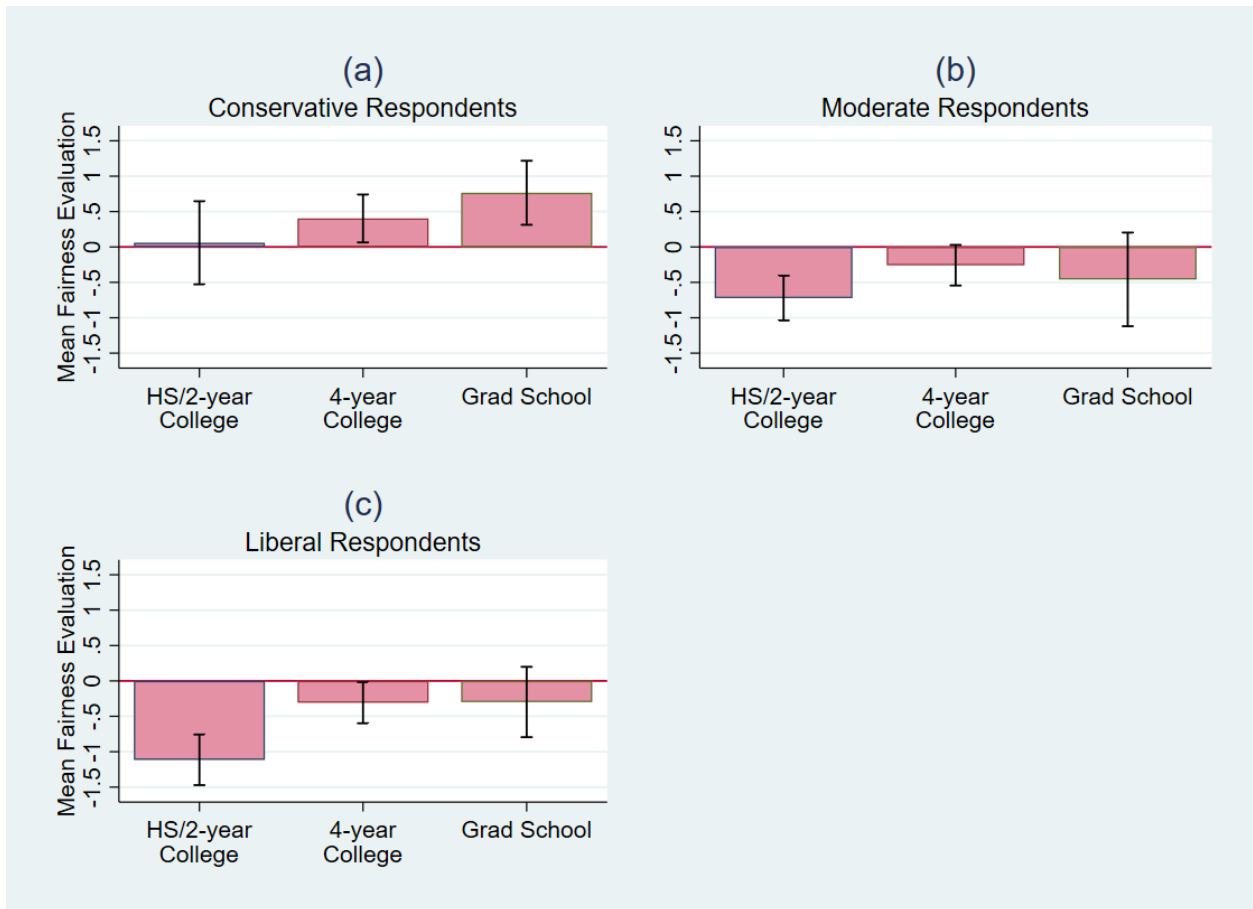
Notes: Fairness is measured on a scale from -3 (“very unfair”) to 3 (“very fair”), where 0 was “neither fair nor unfair.” This figure is based on only Stage 1 observations. The *p*-values below are clustered by respondent.

| | | |
|--|--|---|
| <p>a) Gender: Males vs. Females = 0.037</p> <p>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</p> | <p>c) Respondent Race: White vs. Black = 0.204 Black vs. Other = 0.058 White vs. Other = 0.197</p> <p>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</p> | <p>e) Political party preference Republicans vs. Independents = 0.000 Independents vs. Democrats = 0.961 Democrats vs. Republicans – 0.000</p> <p>f) Political leaning: Conservatives vs. Moderates = 0.000 Moderates vs. Liberals = 0.463 Liberals vs. Conservatives = 0.000</p> |
|--|--|---|

A2.4 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. It shows that this positive association is observed among both conservative and liberal respondents (Figure A2.4.1), and that – 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 A2.4.2). Finally, our result that liberals exhibit a discriminatee race effect and conservatives do not is present within all three education groups in our survey (Figure A2.4.3). Because this education-fairness association is so broadly distributed across political groups and experimental treatments, we conclude that it likely reflects different *set points* for fairness rather than differences in political affiliation or racial attitudes across education groups.

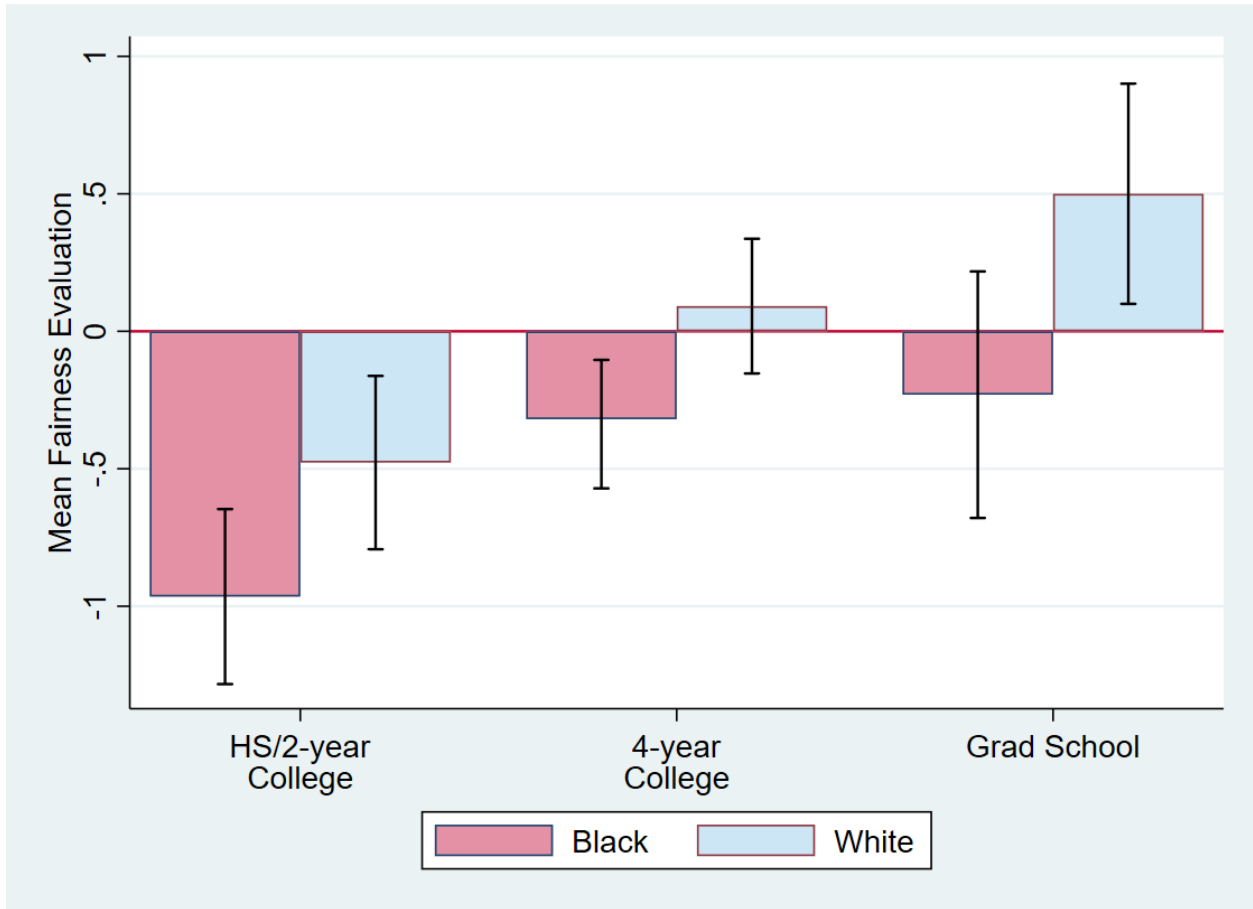
Figure A2.4.1: Mean Fairness Assessments by Education and Political Leaning



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

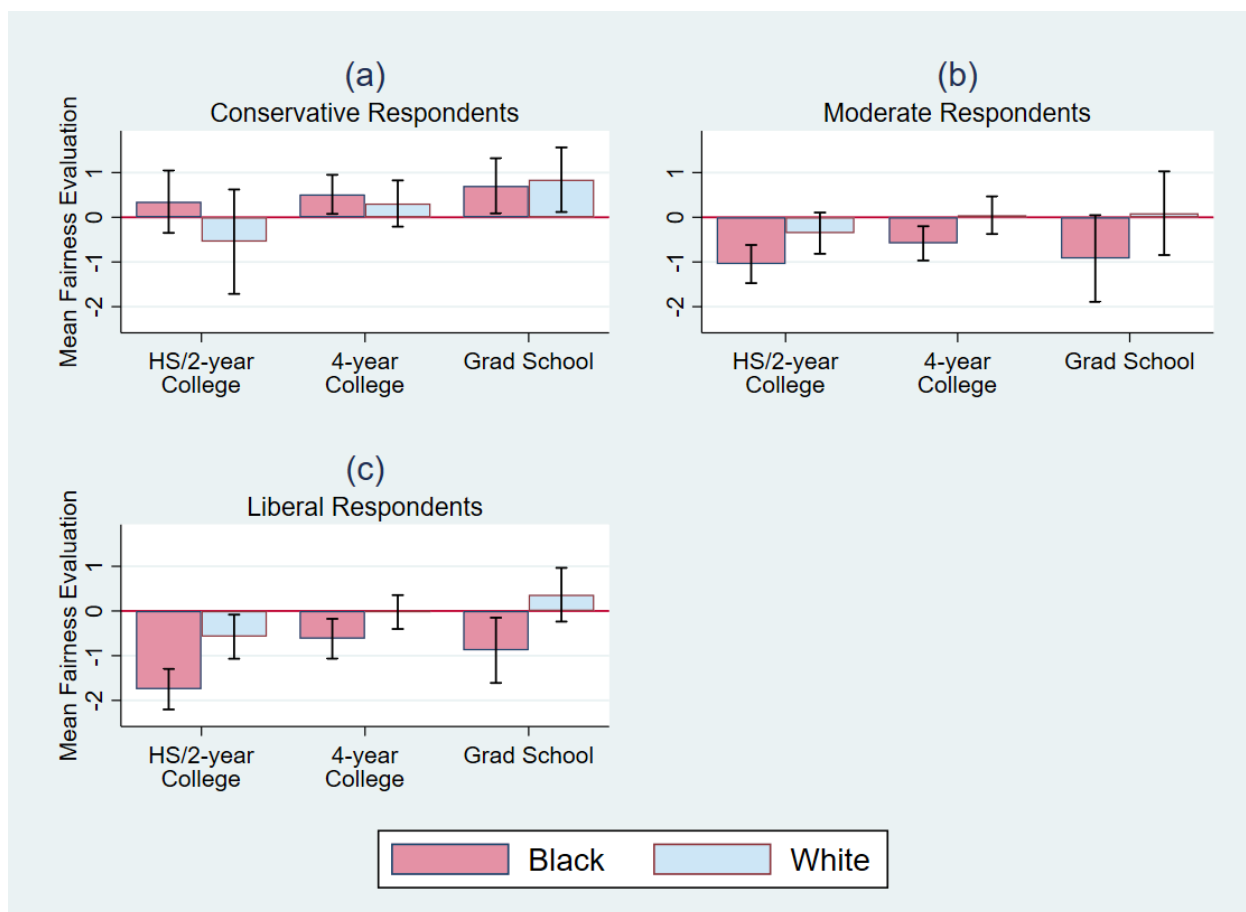
Figure A2.4.2: Discriminatee Race Effects by Education



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

- For HS/2-year 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 A2.4.3: Discriminatee Race Effects by Education and Political Leaning



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

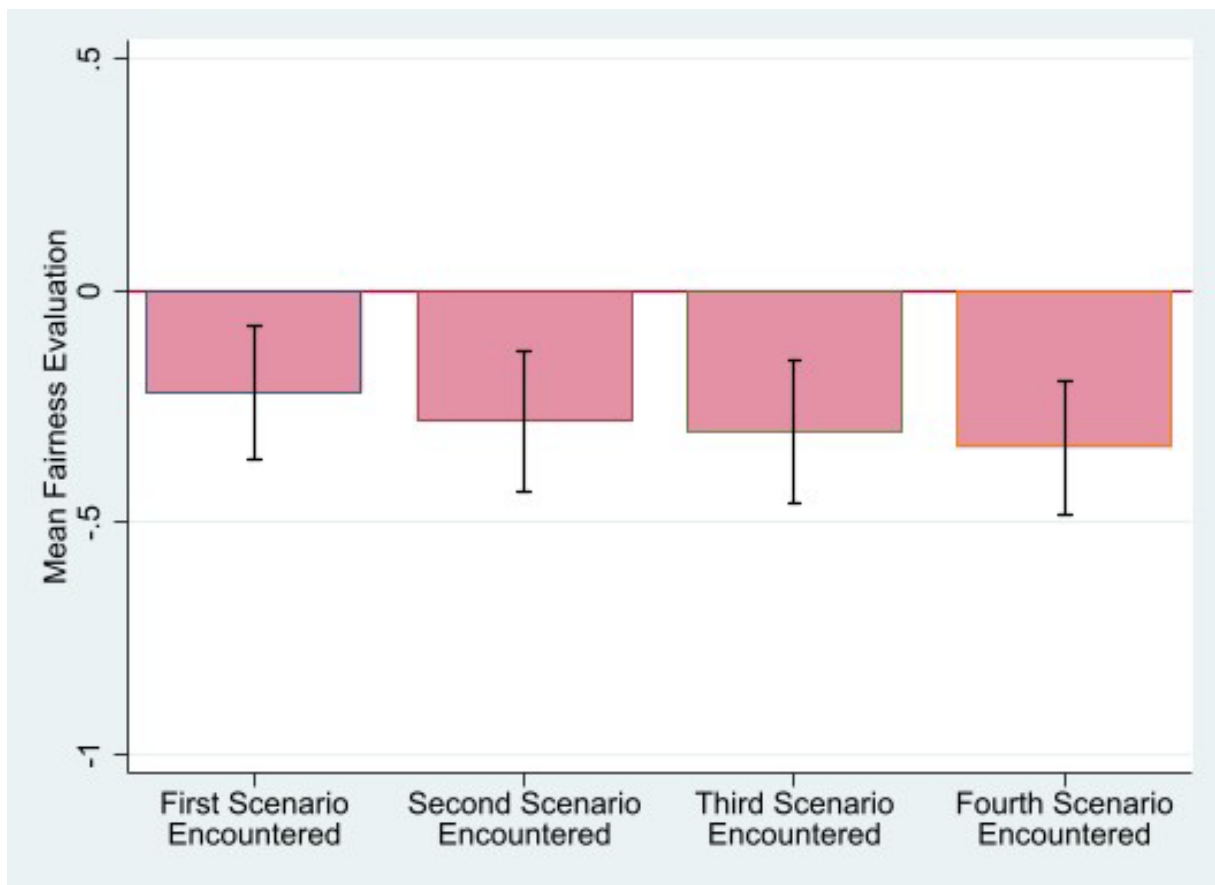
- Conservative respondents:
 - For HS/2-yar College graduates: Black vs. White = 0.154
 - For 4-year College graduates: Black vs. White = 0.544
 - For Graduate School graduates: Black vs. White = 0.765
- Moderate respondents:
 - For HS/2-yar College graduates: Black vs. White = 0.029
 - For 4-year College 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 3: Order Effects

A3.1 Pure Order Effects

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

Figure A3.1.1



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

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

A3.2 Order Effects for the Taste versus Statistical Treatments

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

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

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

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



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

A vs B = 0.834

C vs D = 0.755

A vs C = 0.675

B vs D = 0.314

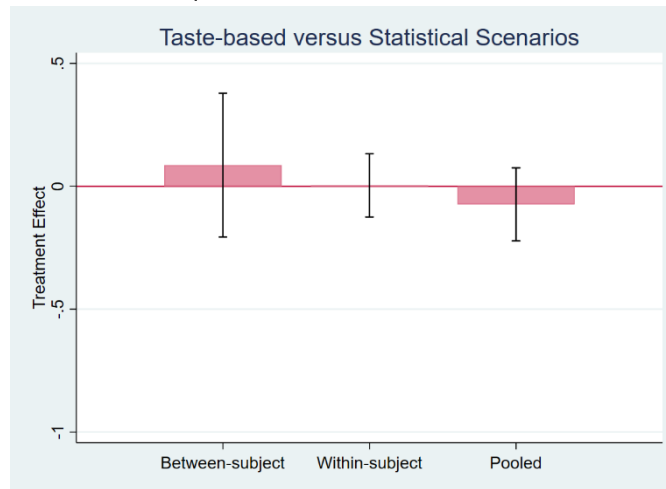
A3.2.2 Ratings Changes of Subjects Who Switched Treatments

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

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

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

Figure A3.2.3: Comparison of Taste Treatment Effect Estimates



Notes:

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

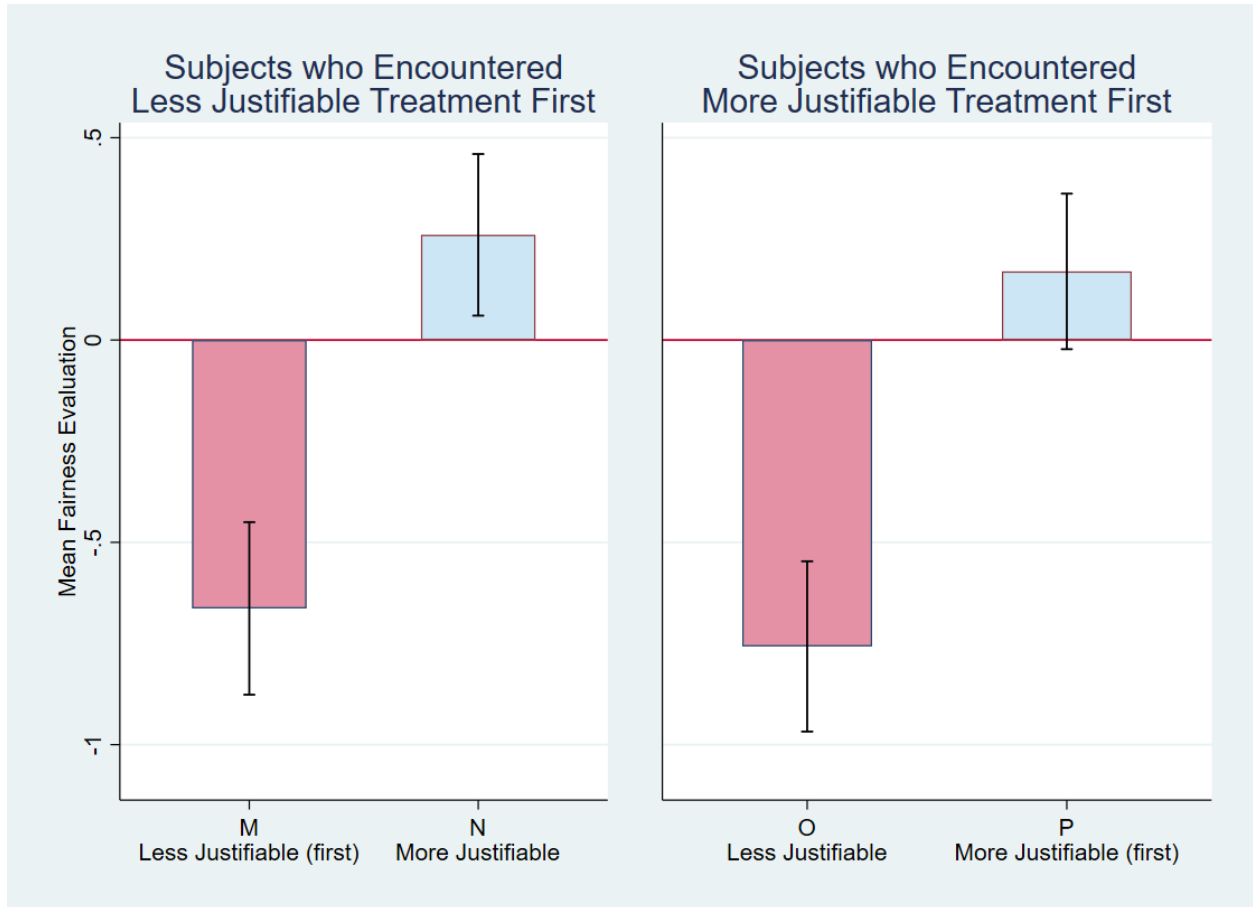
A3.3 Order Effects for the *Less* versus *More* Justifiable Treatments

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

A3.3.1 Justifiability Treatment Variation within Stage 1

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

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



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

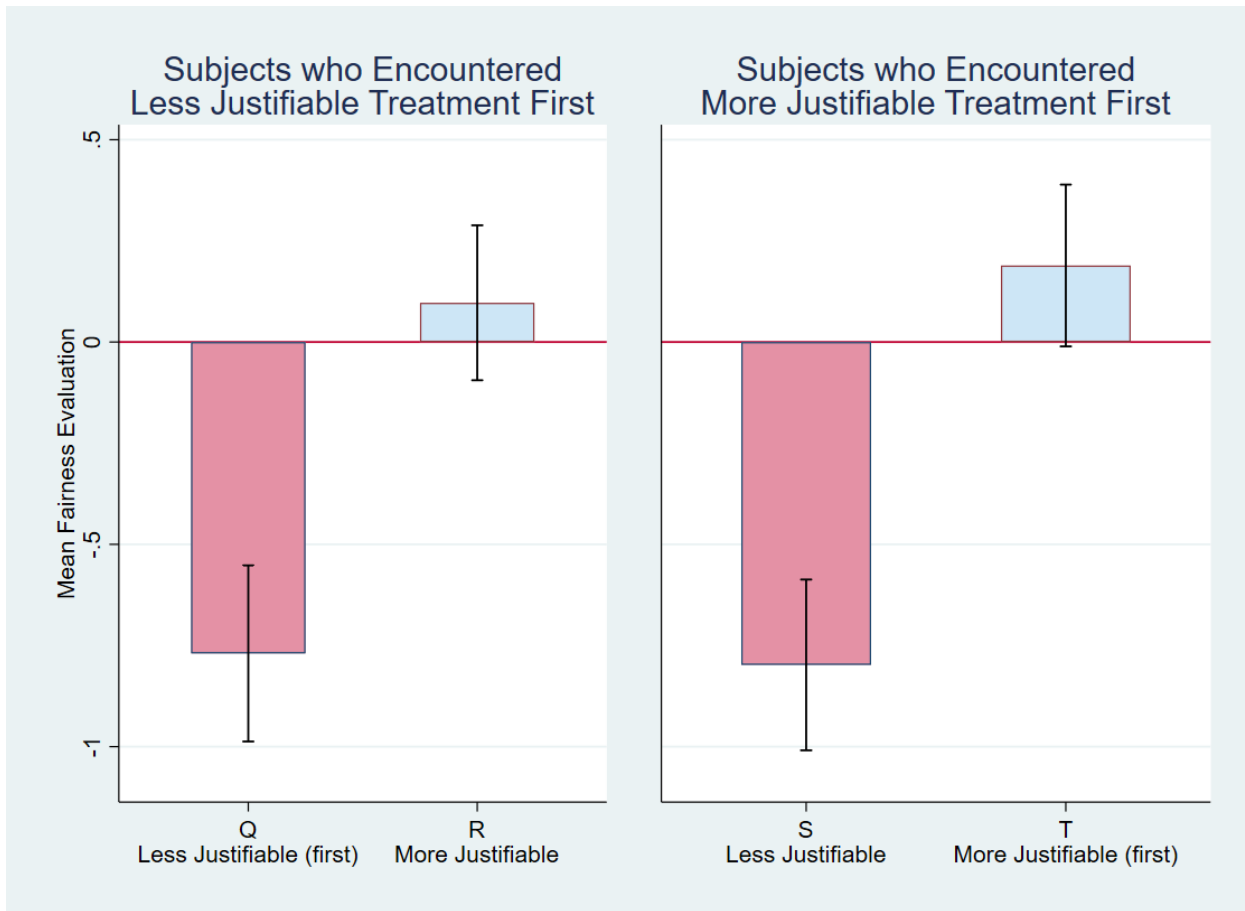
- M vs. N = 0.000
- O vs. P = 0.000
- M vs. O = 0.537
- N vs. P = 0.521

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

A3.3.2 Justifiability Treatment Variation within Stage 2

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

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



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

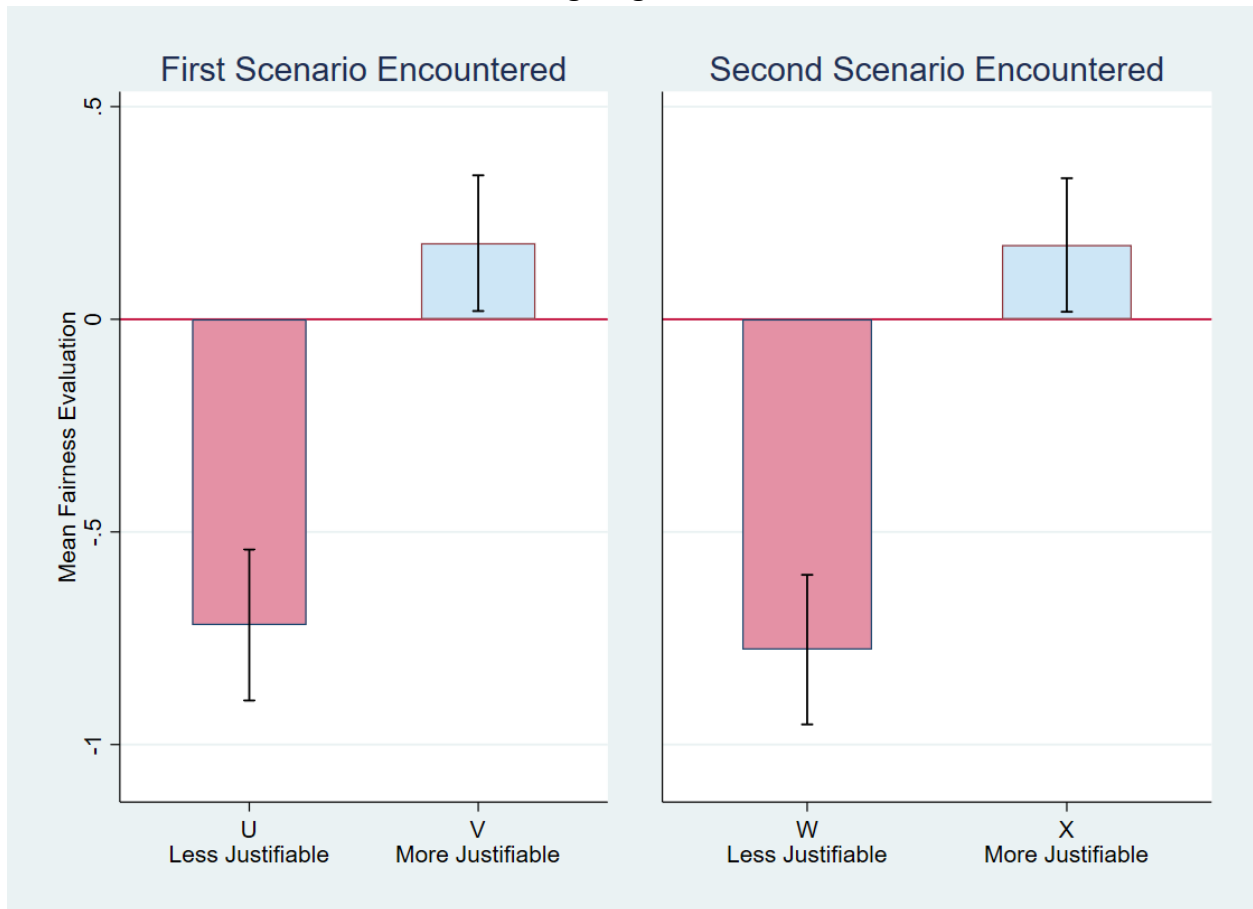
- Q vs. R = 0.000
- R vs. S = 0.000
- Q vs. S = 0.854
- R vs. T = 0.513

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

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

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

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



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

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

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

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

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

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



Notes:

1. The p -values below are clustered by respondent.
 - Between vs. Within-subject = 0.498
 - Within-subject vs. Pooled = 1.00
 - Pooled vs. Between-subject = 0.498
2. Within-subject estimates regress fairness on a treatment indicator plus respondent fixed effects. Between-subject estimates are pure cross-section regressions using data from the first scenario each respondent encountered only. Pooled estimates include all four scenarios each person encountered, without person fixed effects.

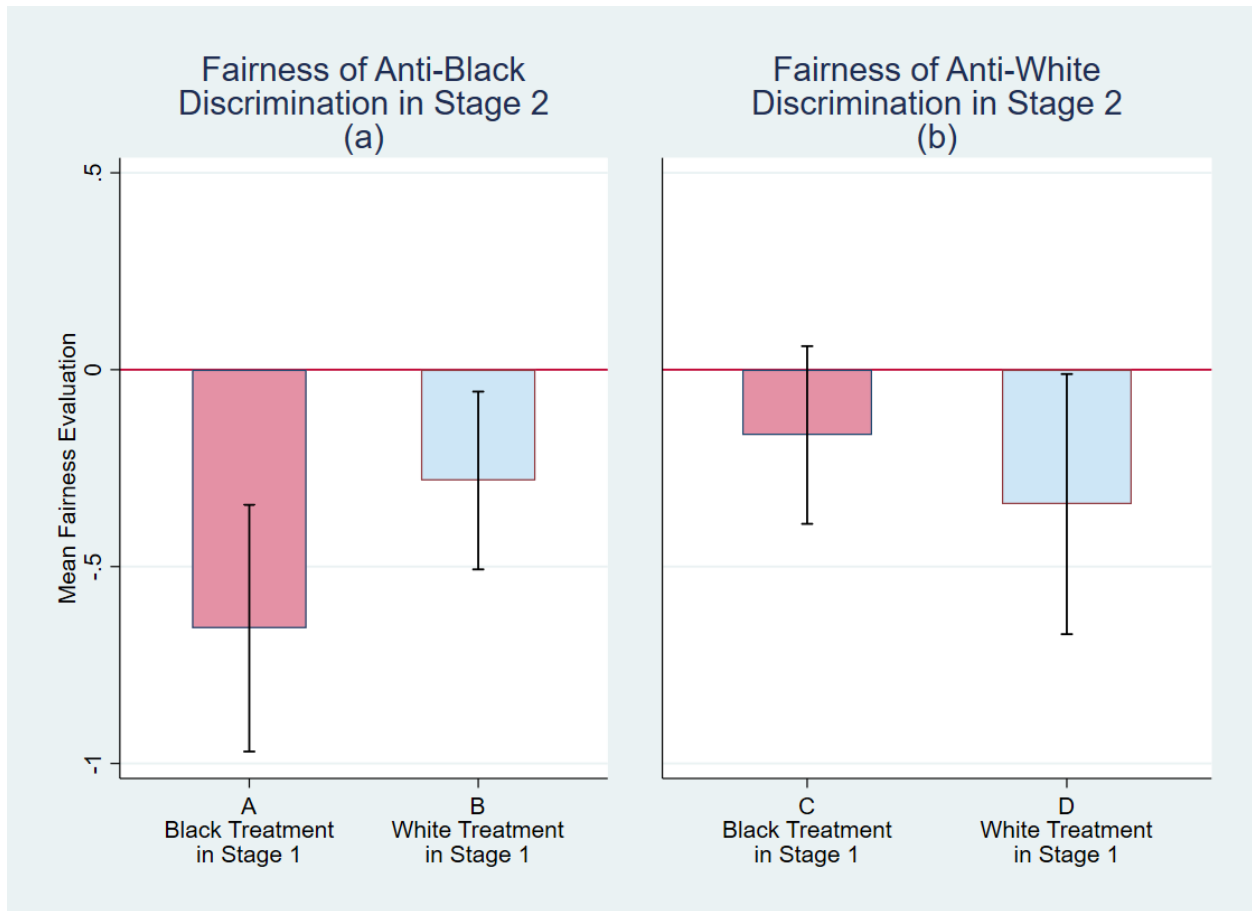
A3.4 Order Effects for the *Race* Treatment

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

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

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

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



p-values:

A vs B = 0.055

C vs D = 0.385

A vs C = 0.012

B vs D = 0.767

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

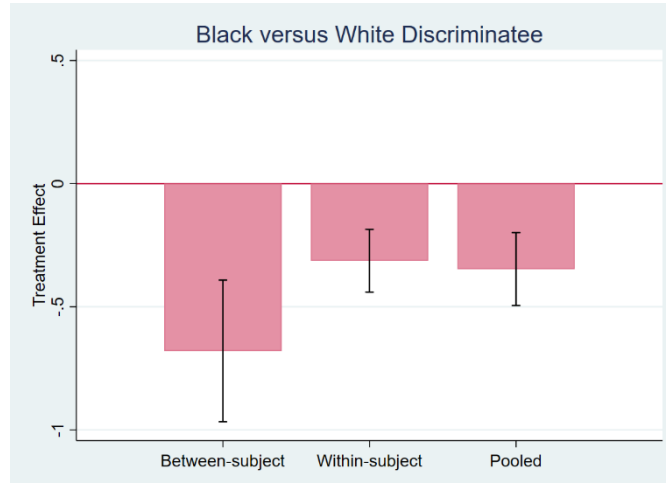
A3.4.2 Ratings Changes of Subjects Who Switched *Race* Treatments

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

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

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

Figure A3.2.3: Comparison of Black Treatment Effect Estimates



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

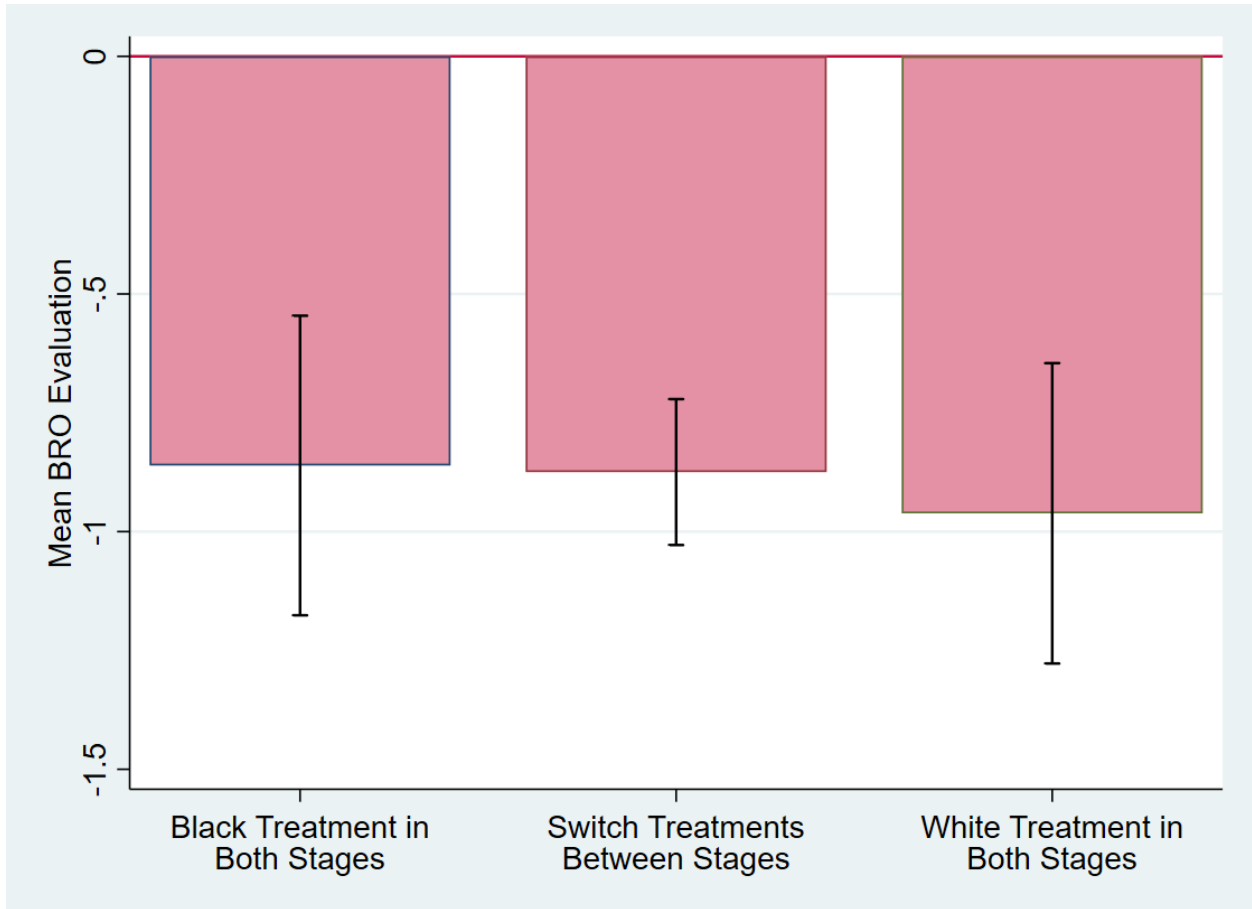
- Between vs. Within-subject = 0.001
- Within-subject vs. Pooled = 0.647
- Pooled vs. Between-subject = 0.007

Appendix 4: Testing for Experimenter Demand Effects using the Relative Opportunities and Political Orientation Questions

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 A4.1 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 A4.2 replicates the analysis for subjects' reported political leaning (on a scale from -3 to +3). Finally, Figure A4.3 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 A4.1: Mean BRO Evaluation Across Respondents' Survey Experience



Notes:

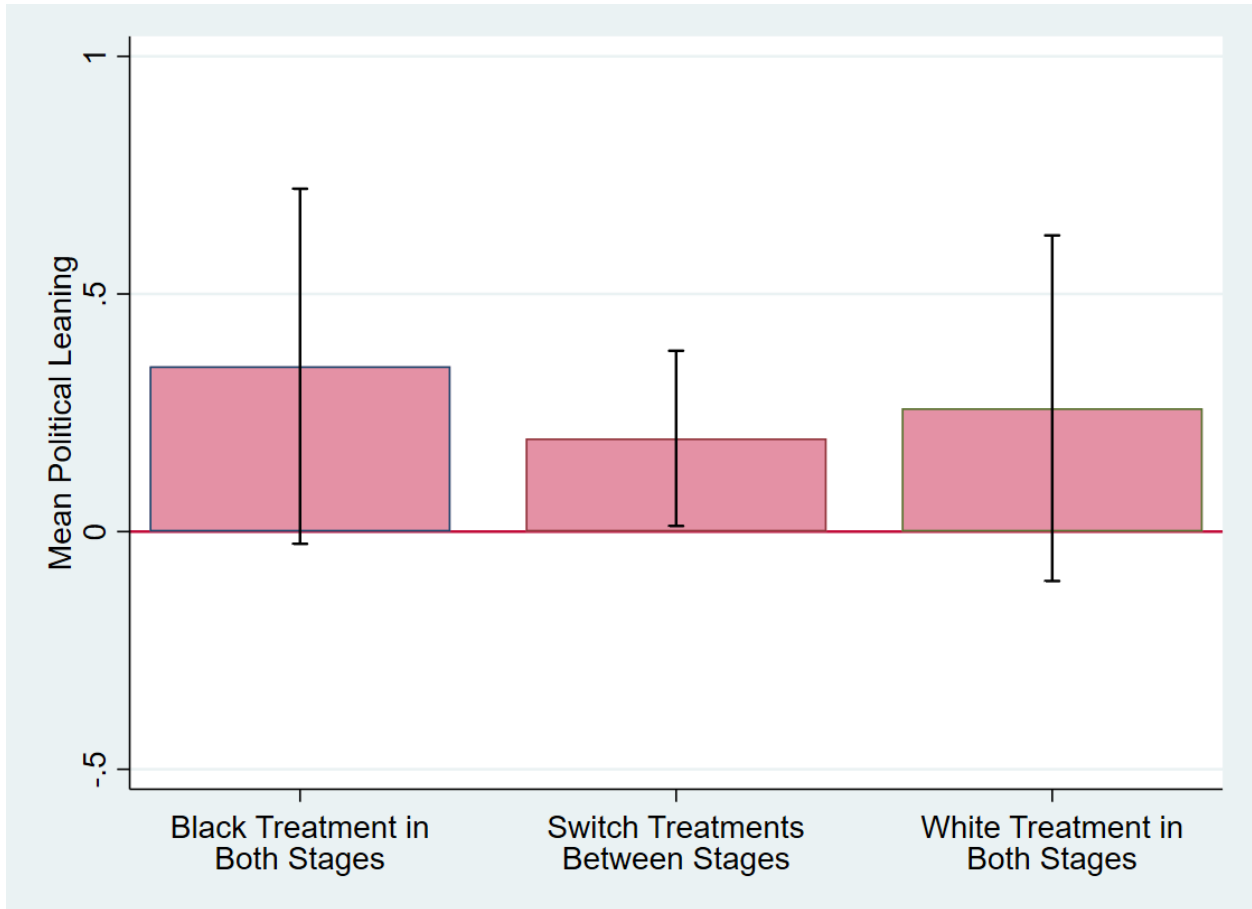
BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more).

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 A4.2: 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).

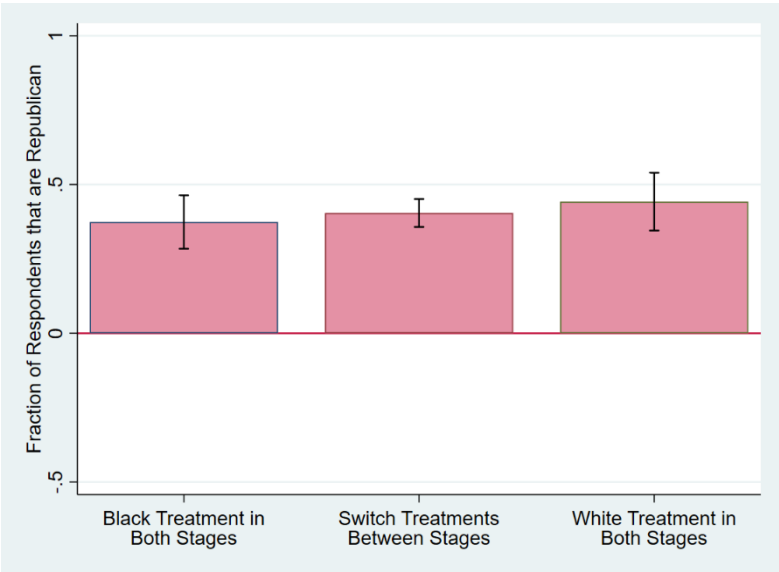
Notes: The p -values below are clustered by respondent.

- 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

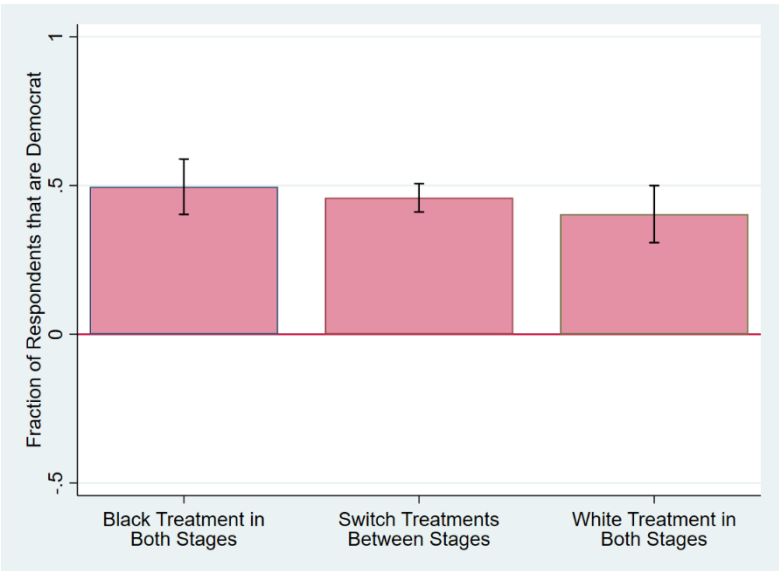
If the respondents modify their reported political leanings to cater to the (inferred) political preferences of the experimenters, we should see a monotonic decrease (shift from liberal towards conservative) from left to right. Such a decrease is not present.

Figure A4.3 Respondents' Survey Experience by Party Preference

(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 5: *Race* Treatment Order Effects by Political Orientation

Appendix 5 studies how the *race* treatment order effects vary with political orientation. It demonstrates that these order effects are absent among conservative respondents, and provides data that allow us to operationalize the 'trade-off' model of moderates' and liberals' ratings changes in Section 5.3 of the paper. Specifically, Appendix 5 replicates Figure A3.4.1 (which showed that subjects' Stage 2 fairness assessments depend on the *race* treatment they encountered in Stage 1) separately for conservative respondents, and for moderate / liberal respondents.

Figure A5.1 shows the mean fairness ratings of conservative respondents in Stage 2, 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, conservatives' 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 conservatives' Stage 2 fairness assessments are unaffected by the treatment they encountered in Stage 1 ($p = .739$ for the Black treatment in Stage 2; $p = .829$ for the White treatment in Stage 2).

Turning to moderates and liberals, Figure A5.2 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 A5.2 also 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. Among these respondents, the Stage 2 fairness difference caused by receiving a different Stage 1 treatment was $-.571$ fairness units ($p = .009$).

Figure A5.1: Race Treatment Order Effects for Conservative Respondents

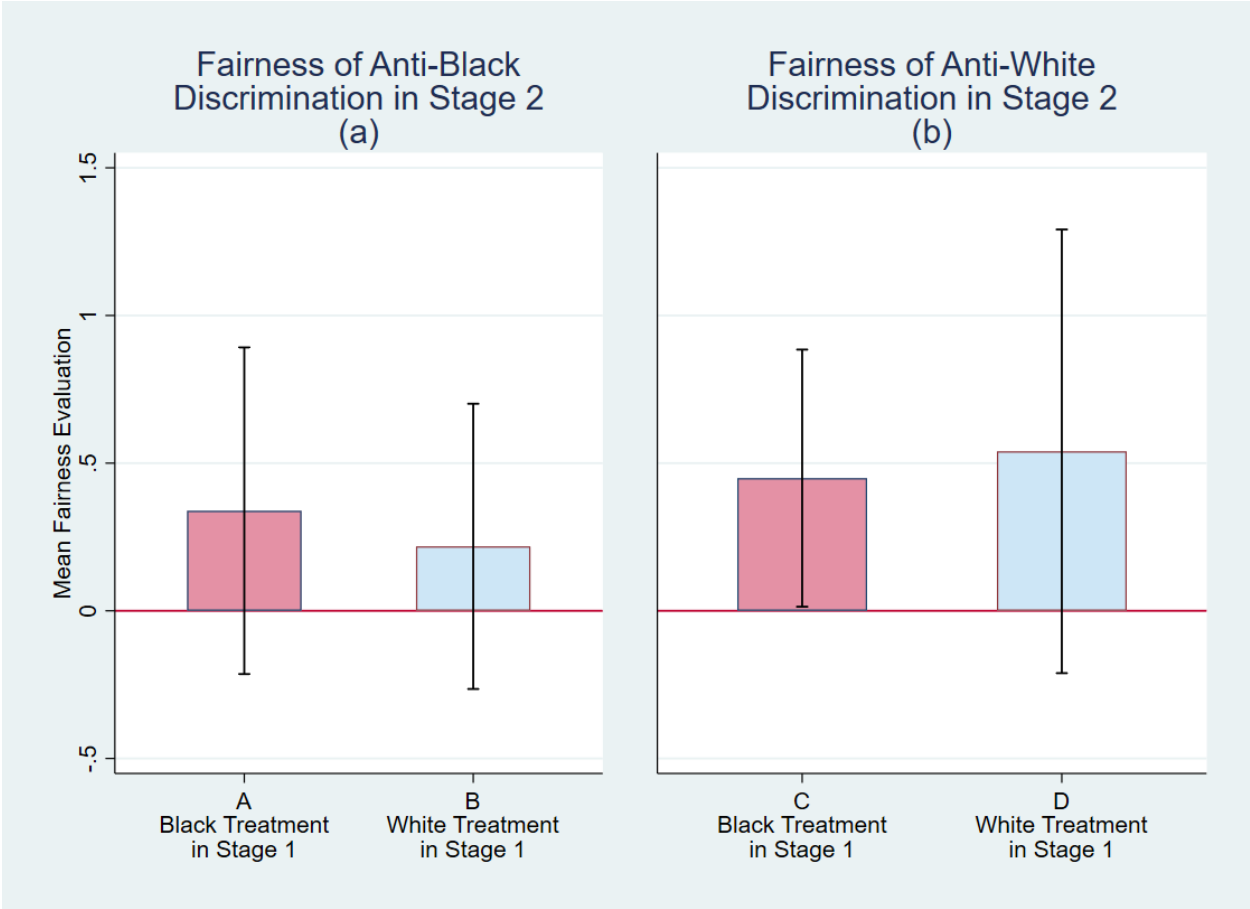


Figure A4.1 shows that conservative respondents’ Stage 2 fairness ratings *do not* depend on the discriminatee race they encountered in Stage 1:

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 A5.2: Race Treatment Order Effects for Moderate and Liberal Respondents

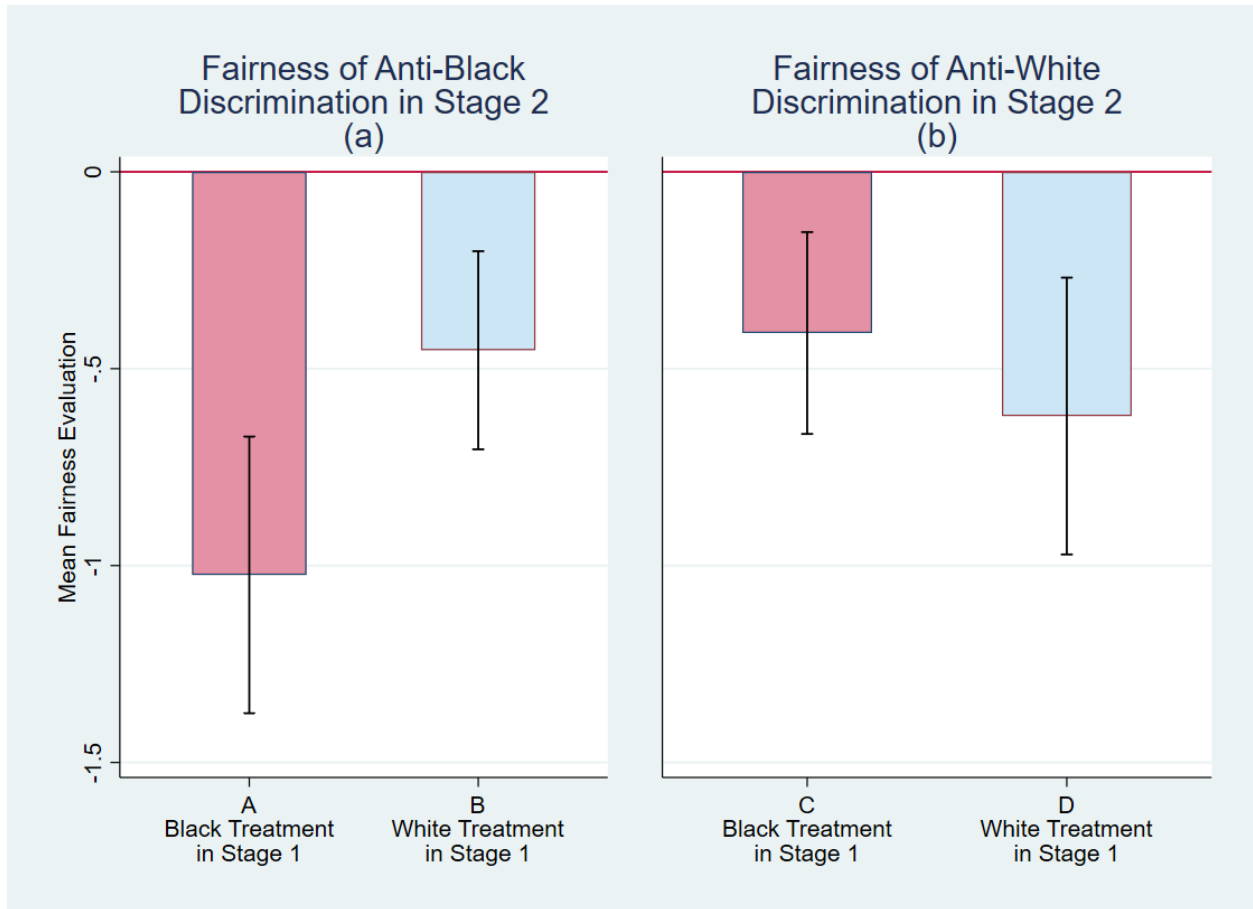


Figure A5.2 shows that moderate and liberal respondents' Stage 2 fairness ratings *do* depend on the discriminatee race they encountered in Stage 1. Specifically, respondents who encountered the Black treatment in Stage 2 rated it as much less fair if they also encountered it in Stage 1 than if they 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.

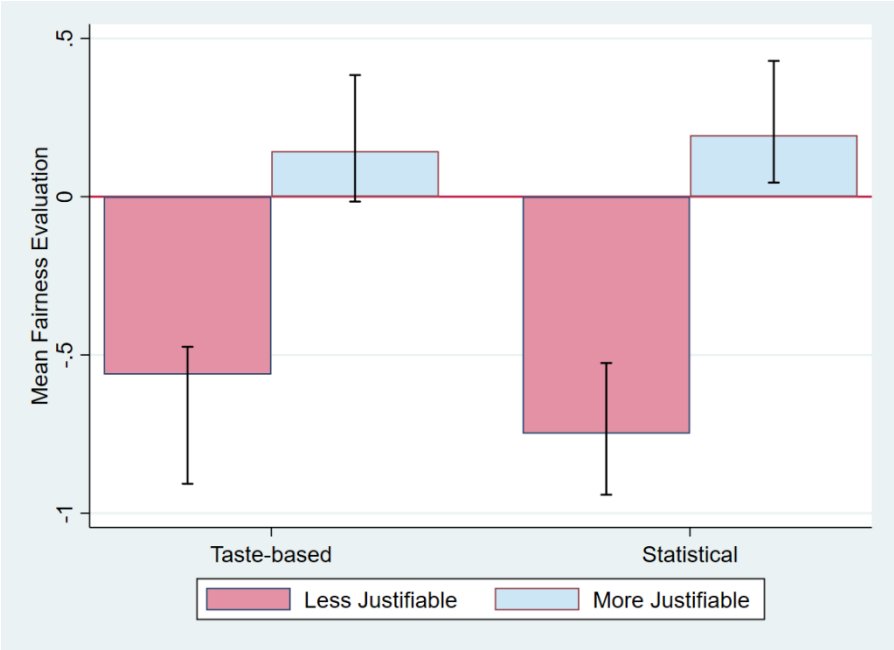
Appendix 6: Additional Robustness Checks

A6.1: Replicating Figures 1 and 2 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 1 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 A6.1.1, are indistinguishable from Figure 1. 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 A6.1.2 repeats this same exercise for Figure 2, which illustrated discriminatee race effects using both scenarios each respondent encountered in Stage 1 of the survey. Figure A6.1.2 shows that the results are extremely similar if we use only information from the very first scenario each respondent encountered in the survey.

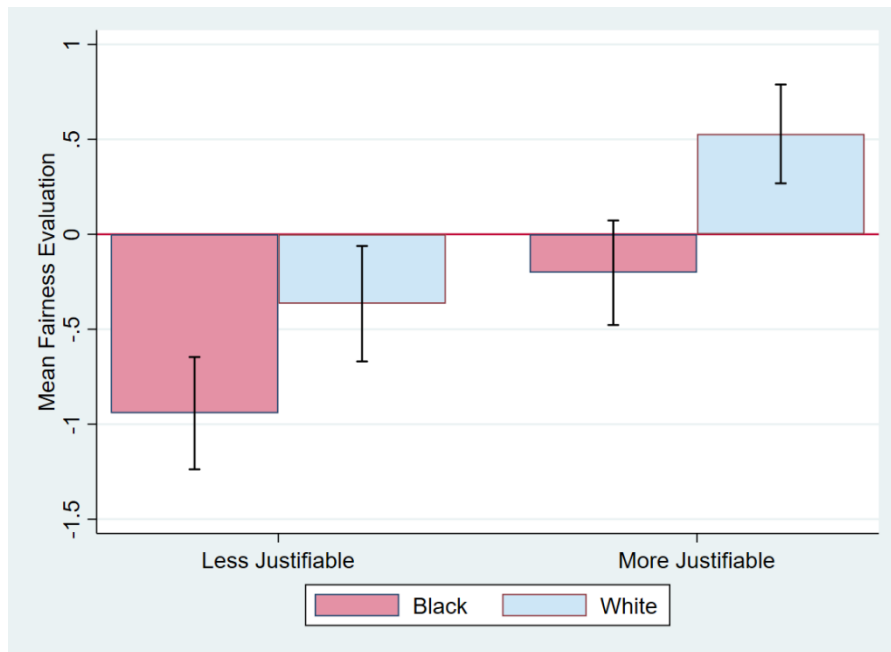
Figure A6.1.1: Fairness Ratings by Type of Discrimination and *Justifiability*: First Scenario Only



Notes: This figure replicates Figure 1 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 A6.1.2: Fairness Ratings by *Justifiability* and Discriminatee Race: First Scenario Only



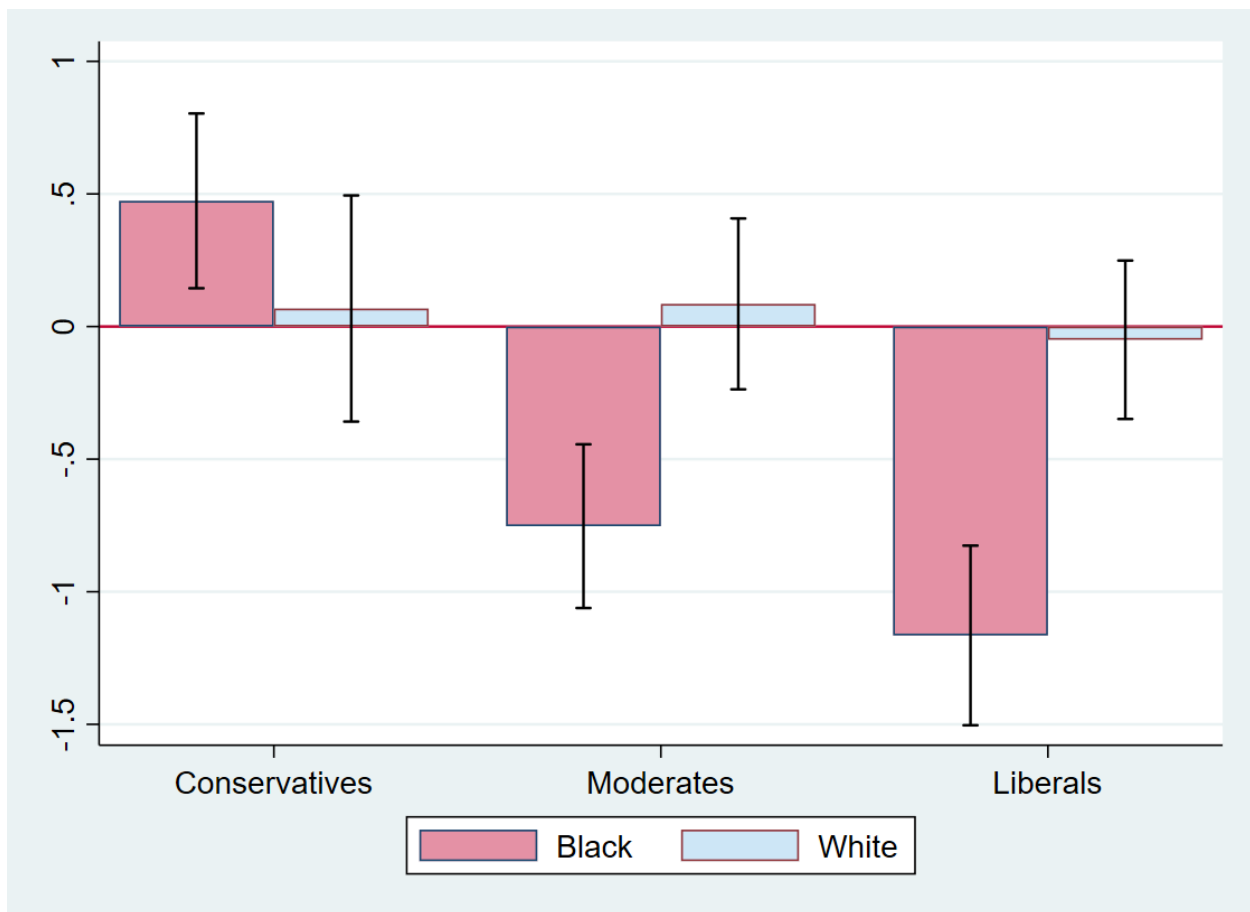
Notes: This figure replicates Figure 2 using only observations from the *first* scenario encountered by respondents in Stage One. Therefore, the *p*-values displayed below are not clustered.

- 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:
 - $p = .5910$

A6.2: Discriminatee Race Effects by Political Orientation for White Respondents Only

To probe the in-group bias hypothesis more deeply, here we replicate Figure 3 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 3 (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 A6.2.1: Discriminatee Race Effects by Political Orientation, White Respondents Only



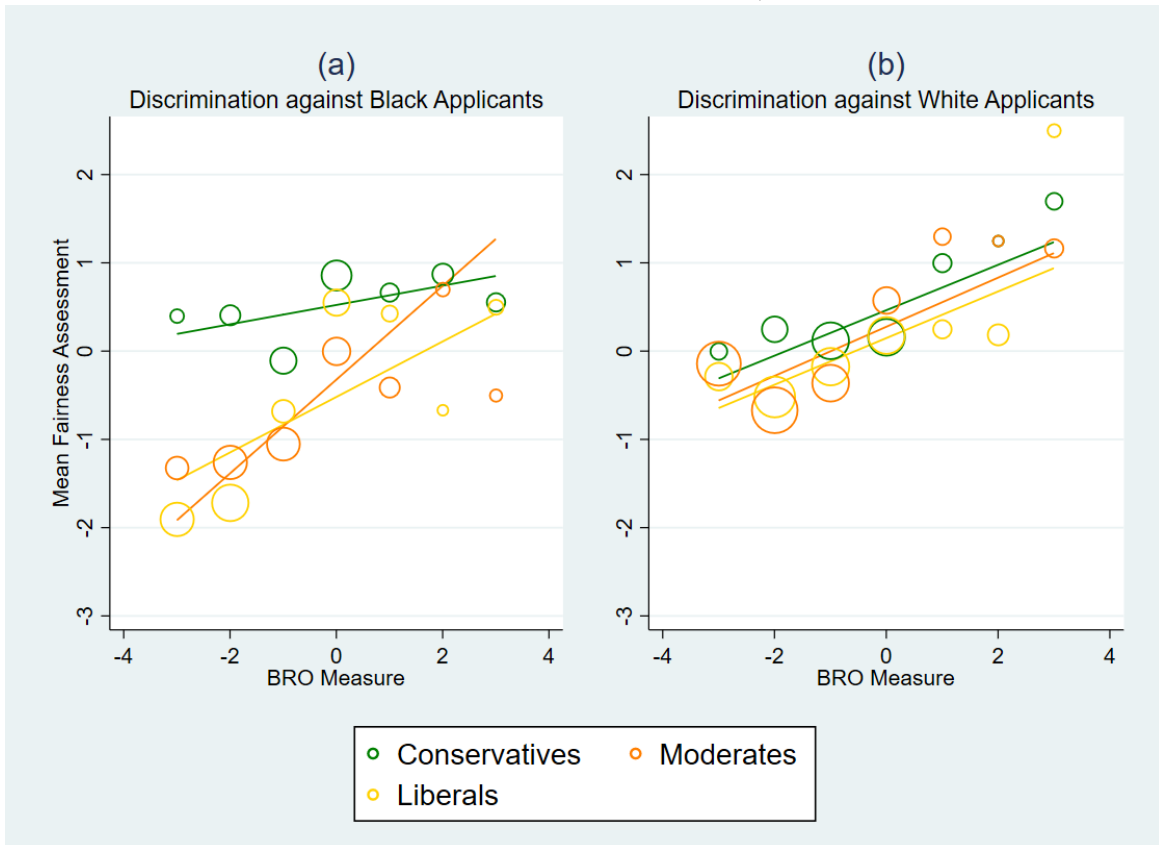
Notes: This figure reproduces Figure 5, but it only reflects the fairness evaluations of White respondents. The p -values below are clustered by respondent.

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

A6.3: Effects of Perceived Relative Opportunities (BRO) on Fairness Ratings, using Three Political Groups

Figure A6.3 replicates Figure 7 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 A6.3: Effects of Perceived Relative Opportunities (BRO) on Fairness Ratings, by Discriminatee Race with Three Political Groups



Notes: This figure reproduces Figure 11, 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 7: Replicating the Main Figures and Tables with ACS Weights

In this Appendix, we replicate Figures 1-7 and Table 1 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 A7.1 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 A7.1 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.

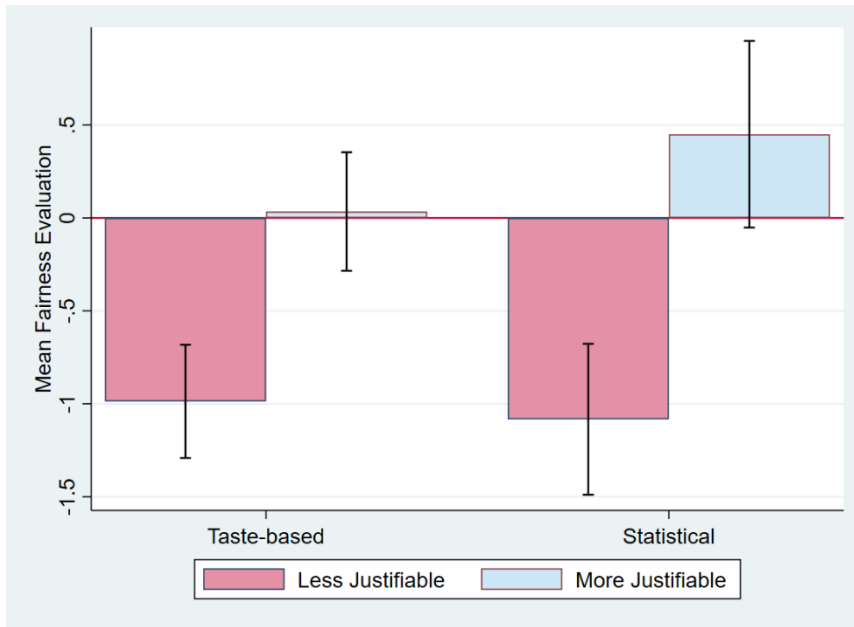
The remaining exhibits in this Appendix replicate Figures 1-7 and Table 1 using these weights. All the main patterns discussed in the paper are also present here, with one small exception: the weak positive association between BRO and the fairness of discrimination among conservative respondents in Figure 7(a) becomes somewhat stronger and statistically significant. Similar to Figure 7, however, the slope for conservatives remains much lower than the slope for moderates / liberals.

Table A7.1: 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 A7.1: Fairness Ratings by Type of Discrimination and *Justifiability* (replicates Figure 1)



p-values:

Less- versus more justifiable treatments:

Overall: $p=.000$
 Within taste-based: $p=.000$
 Within statistical: $p=.000$

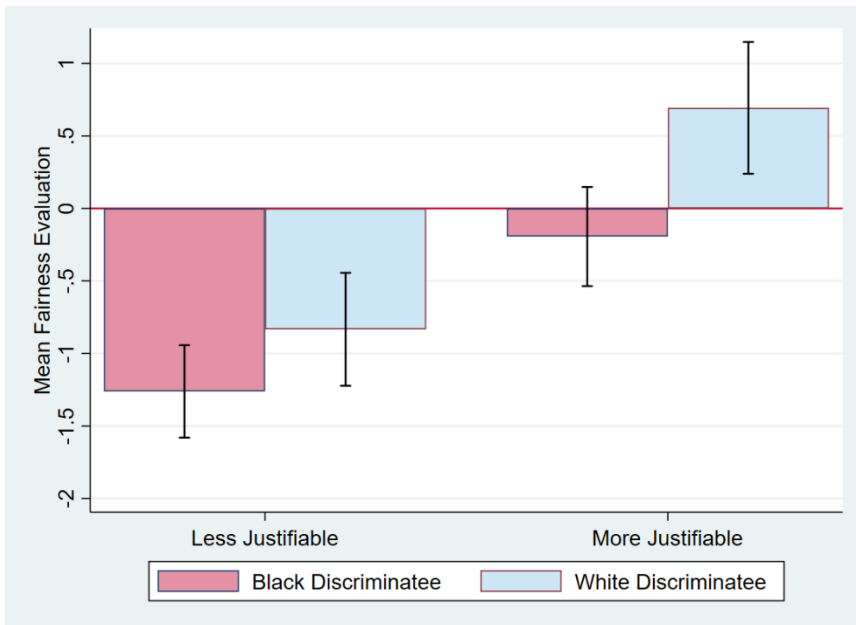
Taste versus Statistical Discrimination:

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 A7.2: Fairness by *Justifiability* and Discriminatee Race (replicates Figure 2)



p-values:

Black versus White Treatment:

Overall: $p=.003$
 Within Less-Justifiable: $p=.095$
 Within More-Justifiable: $p=.002$

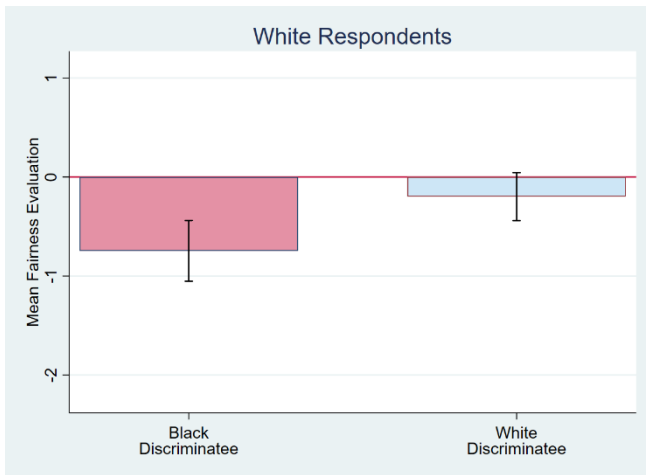
Less versus More Justifiable Treatment:

Overall: $p=.000$
 Within Black Discriminatees: $p=.000$
 Within White Discriminatees: $p=.000$

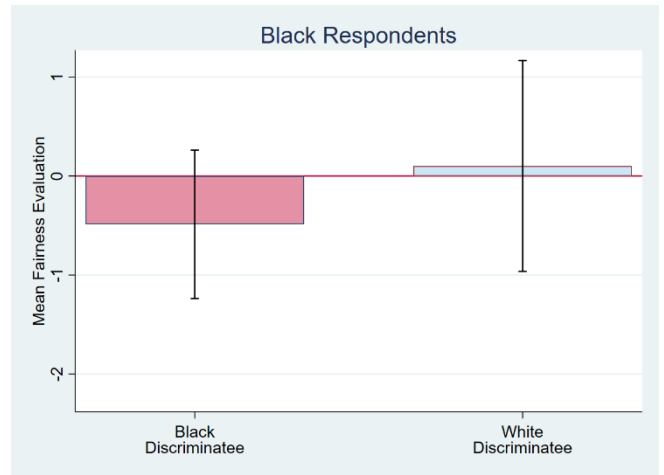
Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 1.068 units less fair. Within White Discriminatees, less-justifiable scenarios are 1.527 units less fair. A test for equality of the Less versus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

Figure A7.3: Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 3)

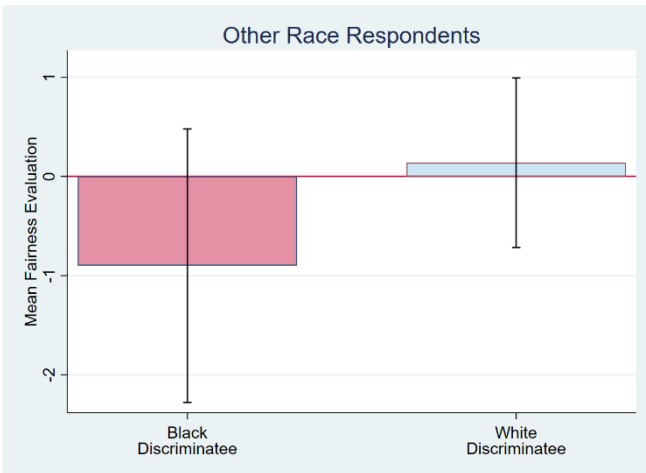
(a)



(b)



(c)



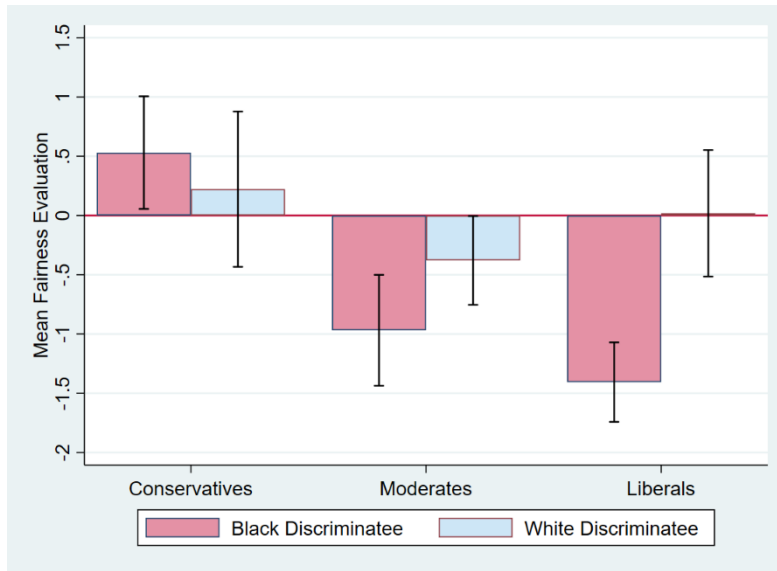
p-values:

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 A7.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Figure 4)



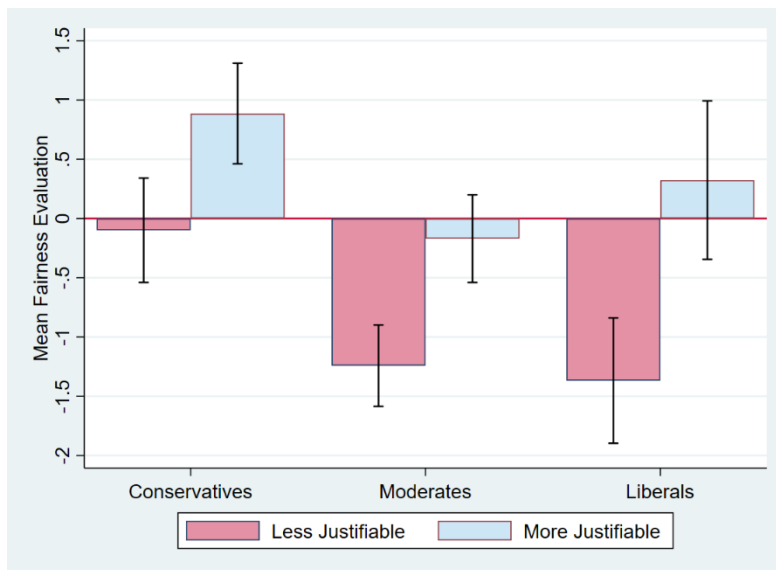
p-values:

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 A7.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent’s Political Leaning (replicates Figure 5)



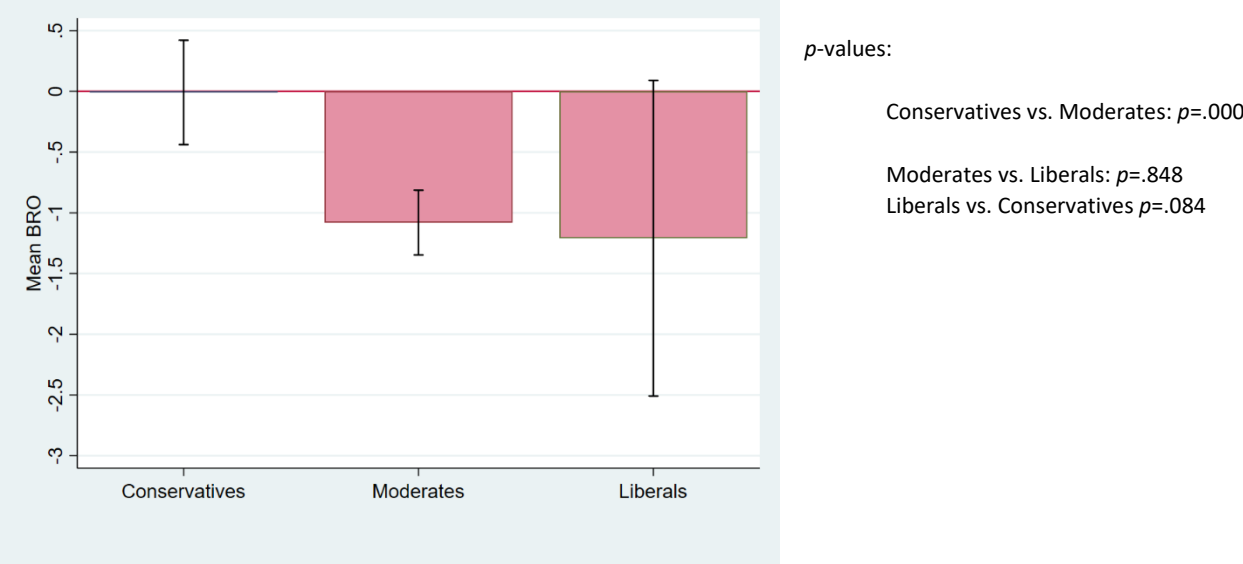
p-values:

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 $p = .153$.

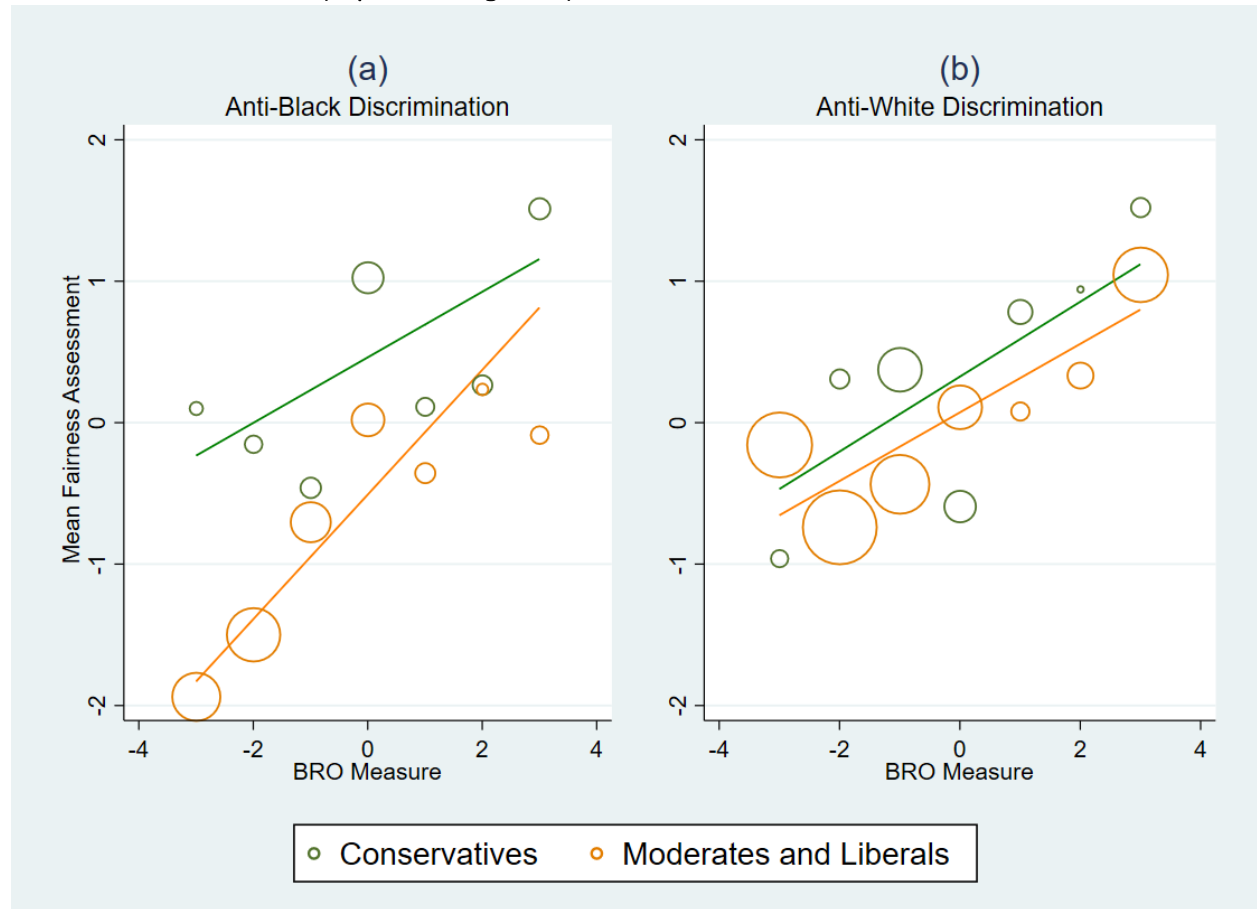
Figure A7.6: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 6)



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 A7.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 7)



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, $p = .000$

Table A7.2: Estimating the Combined Effects of Utilitarian and Rules-Based Determinants of Fairness (replicates Table 1)

| | All (1) | All (2) | Conservatives (3) | Conservatives (4) | Moderates and Liberals (5) | Moderates and Liberals (6) |
|---|----------------------|----------------------|----------------------|----------------------|----------------------------------|----------------------------------|
| Less justifiable | -1.305*** (0.166) | -1.527*** (0.270) | -0.985*** (0.184) | -0.838*** (0.231) | -1.396*** (0.202) | -1.689*** (0.315) |
| Black discriminatee | -0.658*** (0.224) | -0.887*** (0.289) | 0.308 (0.407) | 0.441 (0.431) | -1.017*** (0.253) | -1.334*** (0.331) |
| Less justifiable × Black discriminatee | | 0.459 (0.311) | | -0.265 (0.357) | | 0.633* (0.365) |
| Constant | 0.582*** (0.201) | 0.693*** (0.231) | 0.714** (0.334) | 0.641* (0.332) | 0.559** (0.243) | 0.706** (0.273) |
| Observations | 1,284 | 1,284 | 340 | 340 | 944 | 944 |
| R-squared | 0.076 | 0.076 | 0.055 | 0.055 | 0.110 | 0.110 |

Notes: Regression results are based on Stage 1 fairness assessments only. 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 respondent.

Appendix 8: Replicating the Main Figures and Tables with GSS Weights

In this Appendix, we replicate Figures 1-7 and Table 1 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 A8.1 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.⁴⁸ 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 A8.1 mimics the GSS sample perfectly.⁴⁹

The remaining exhibits in this Appendix replicate Figures 1-7 and Table 1 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 7 does not occur here, suggesting that the unusual political mix of MTurkers is not responsible for *any* of the main results in the paper.

⁴⁸ 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 A2.2 for additional details.

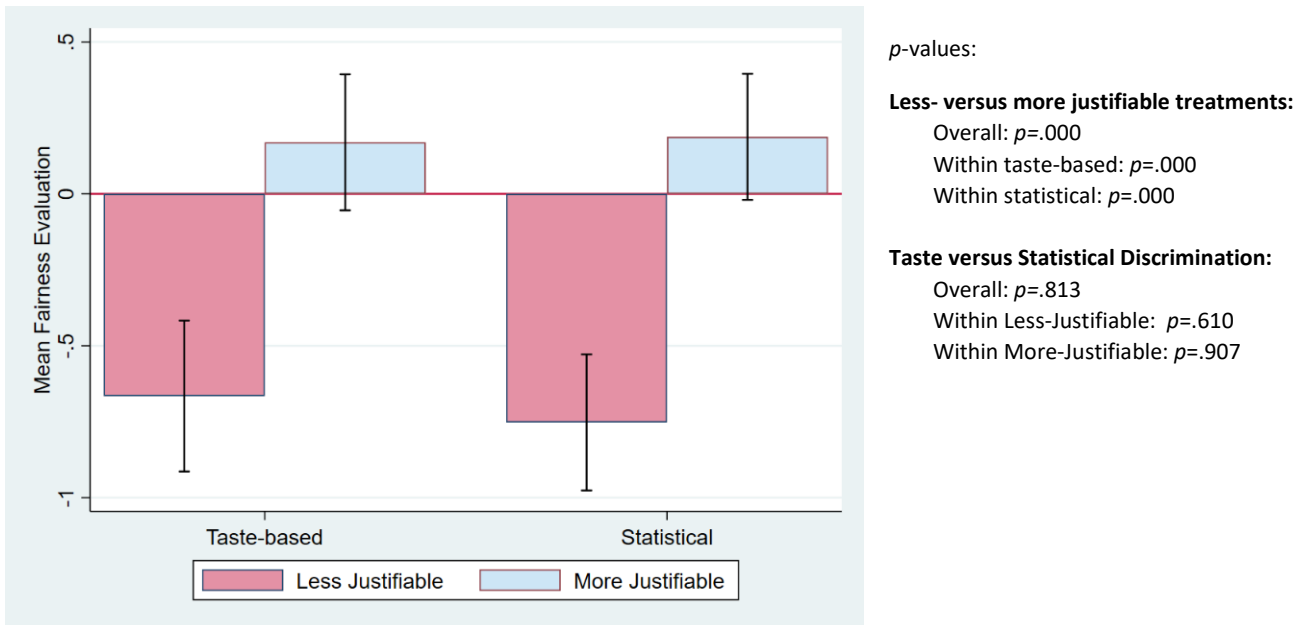
⁴⁹ Because of the small size of the MTurk and GSS samples, we did not re-weight our MTurk sample to mimic GSS demographic characteristics; attempts to do this yielded highly extreme and imprecise weights.

Table A8.1: 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 |

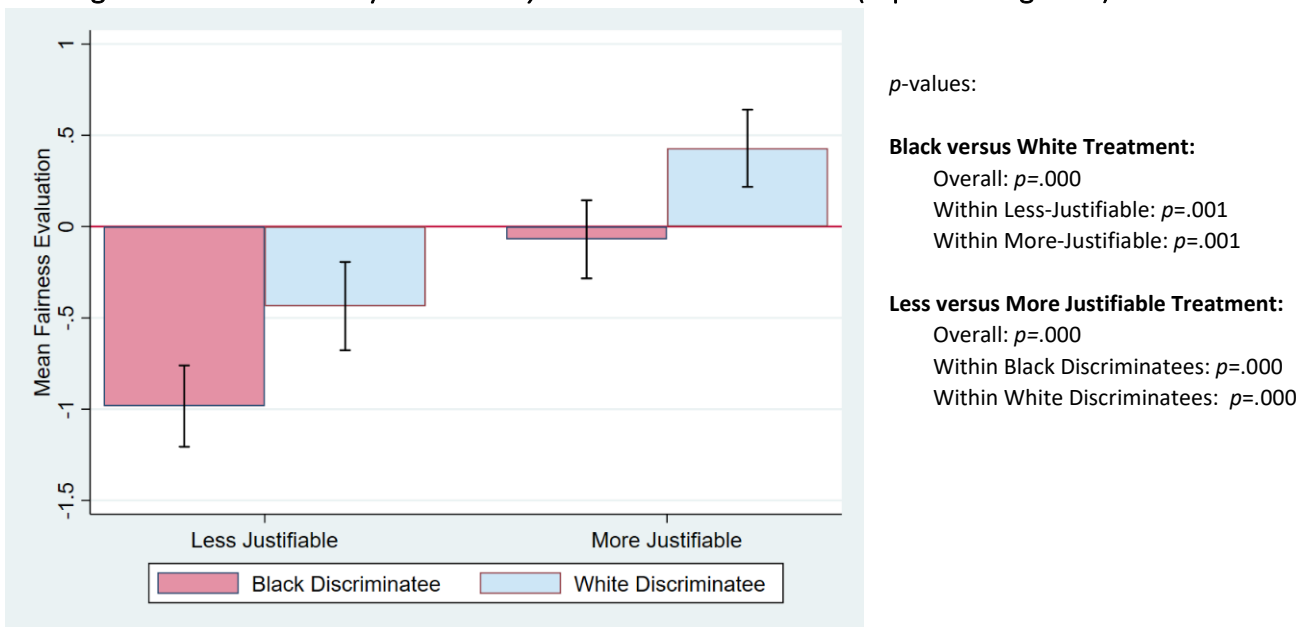
Notes: Column 1 contains the percentage of respondents across various demographic characteristics within the MTurk sample. Column 2 contains these percentages for the 2019 American Community Survey (ACS) sample for comparison.

Figure A8.1: Fairness Ratings by Type of Discrimination and *Justifiability* (replicates Figure 1)



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

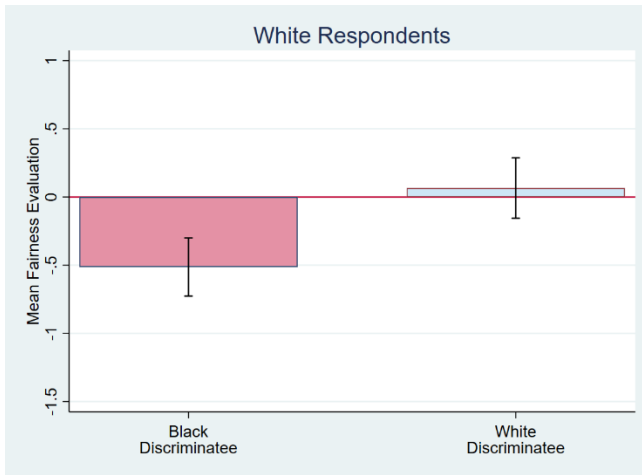
Figure A8.2: Fairness by *Justifiability* and Discriminatee Race (replicates Figure 2)



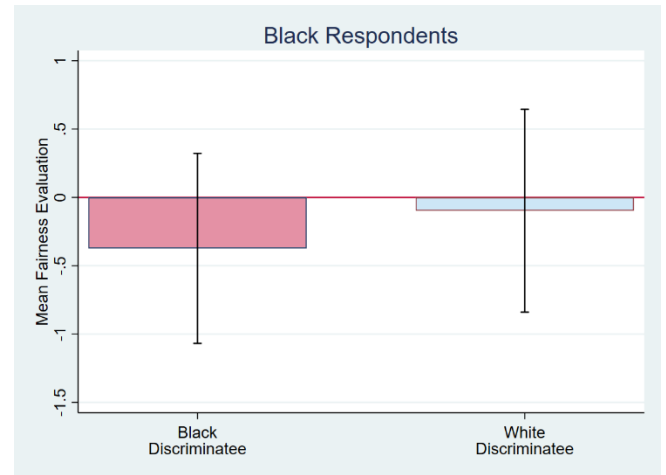
Note: This figure is based on only Stage 1 observations. All p-values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.914 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.865 units less fair. A test for equality of the Less versus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

Figure A8.3: Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 3)

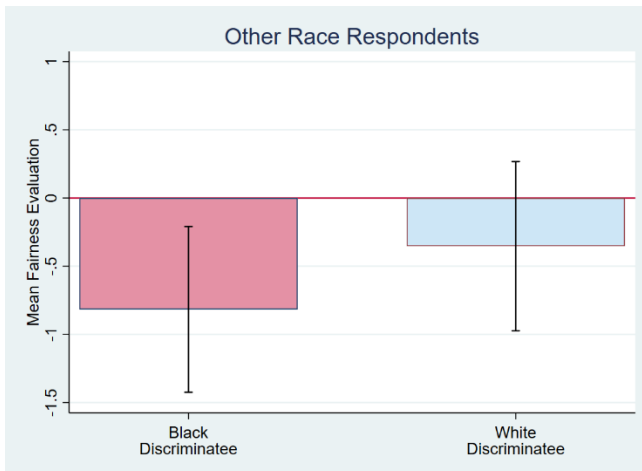
(a)



(b)



(c)



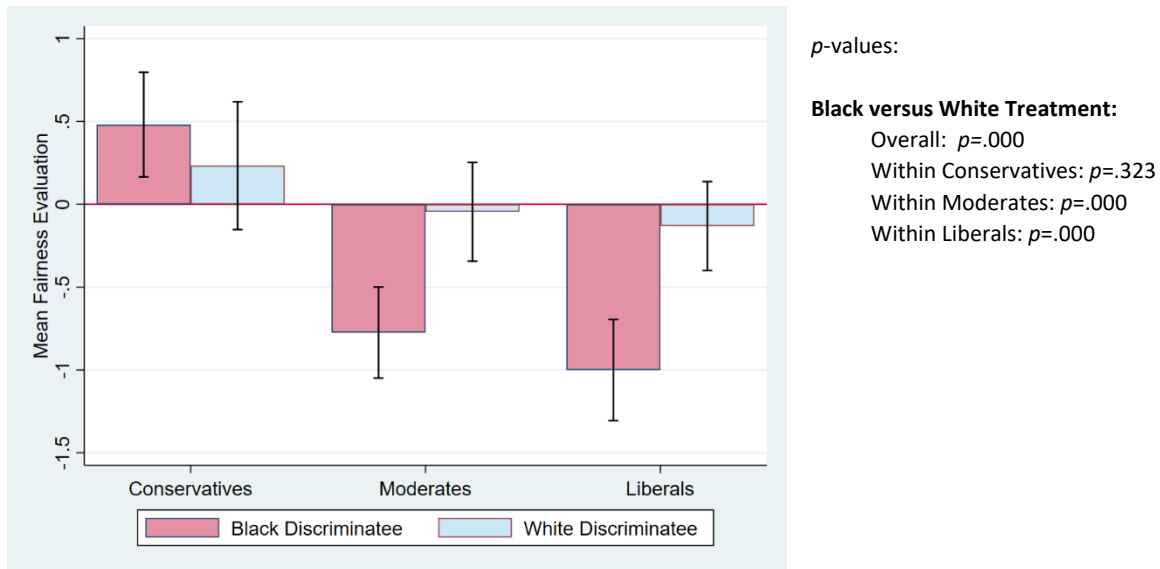
p-values:

Black versus White Treatment:

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

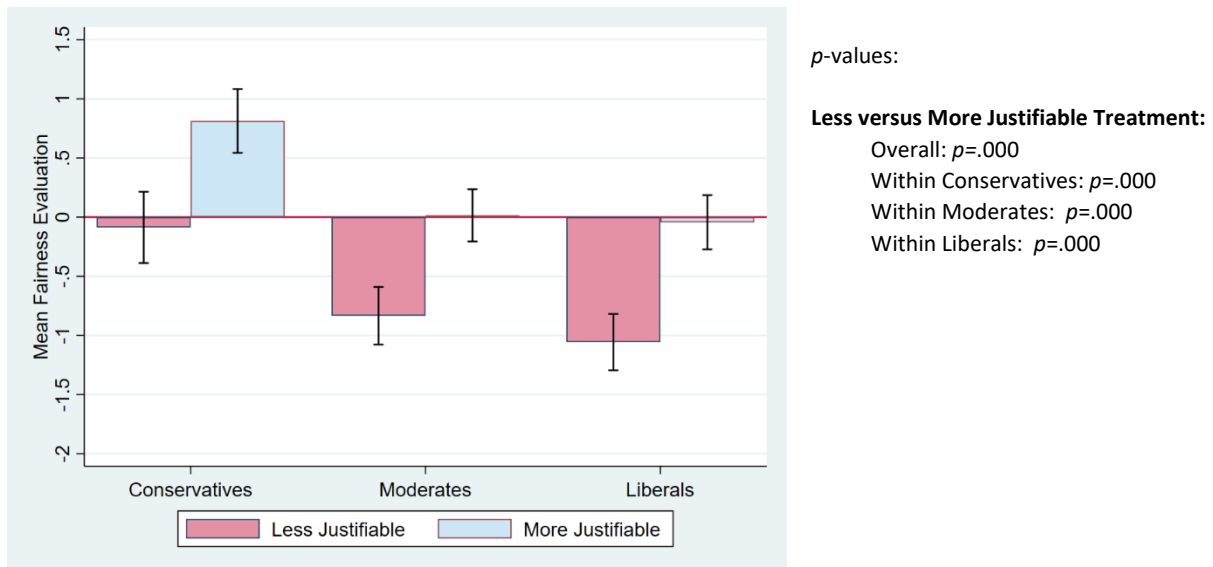
Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e. the Black treatment) across all three racial groups yields $p = .827$.

Figure A8.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Figure 4)



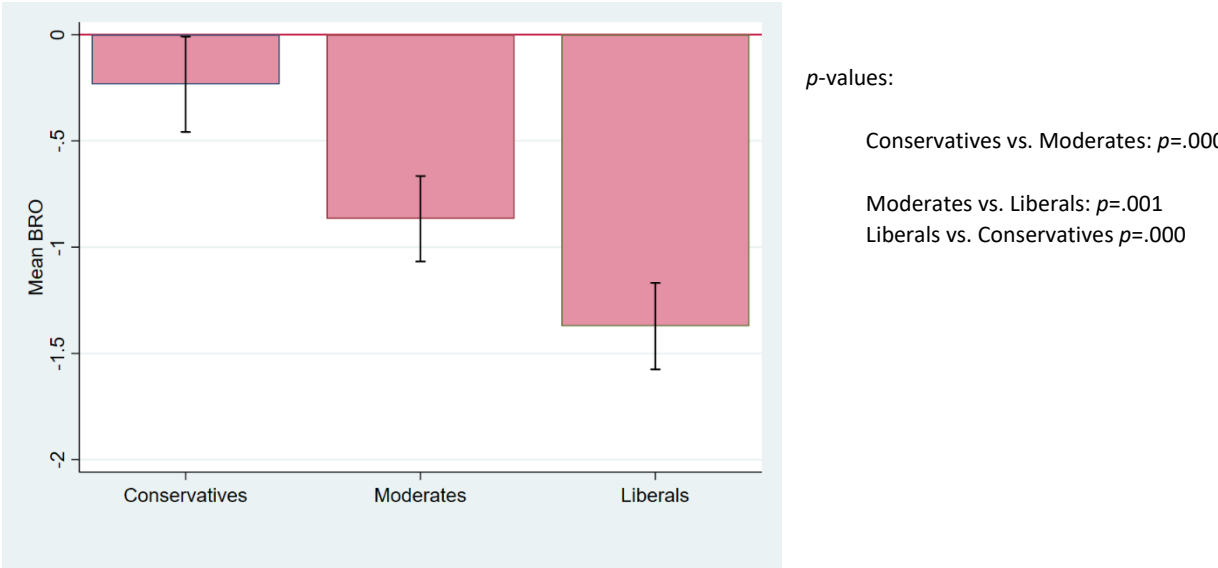
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields $p = .628$. A test for equality between conservatives and (moderates +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

Figure A8.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent's Political Leaning (replicates Figure 5)



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 $p = .541$.

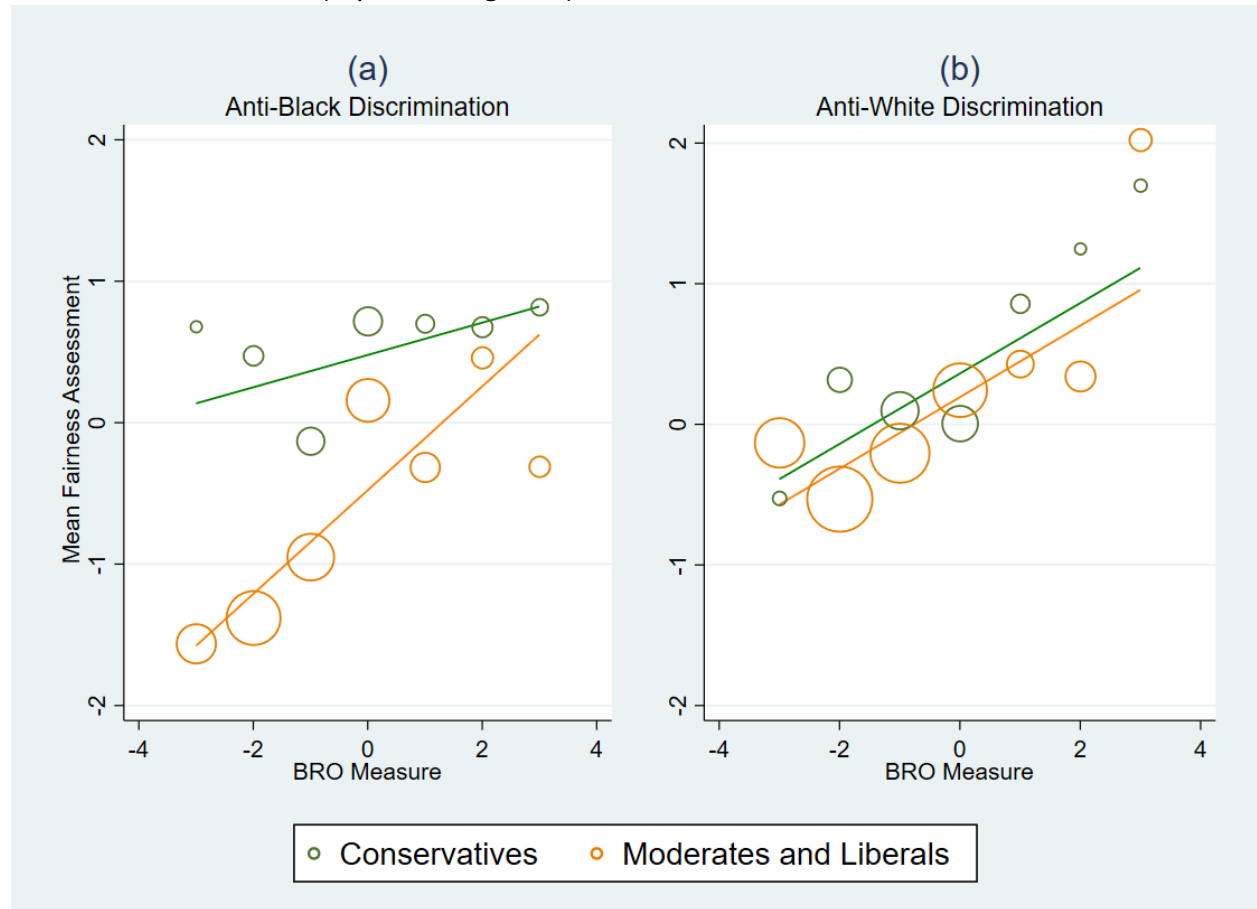
Figure A8.6: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 6)



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 A8.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 7)



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, $p = .001$

Table A8.2: Estimating the Combined Effects of Utilitarian and Rules-Based Determinants of Fairness (replicates Table 1)

| | All (1) | All (2) | Conservatives (3) | Conservatives (4) | Moderates and Liberals (5) | Moderates and Liberals (6) |
|---|----------------------|----------------------|----------------------|----------------------|----------------------------------|----------------------------------|
| Less justifiable | -1.305*** (0.166) | -0.865*** (0.107) | -0.899*** (0.149) | -0.841*** (0.208) | -0.887*** (0.0855) | -0.871*** (0.125) |
| Black discriminatee | -0.658*** (0.224) | -0.499*** (0.153) | 0.248 (0.251) | 0.304 (0.276) | -0.759*** (0.164) | -0.744*** (0.176) |
| Less justifiable × Black discriminatee | | -0.0486 (0.148) | | -0.111 (0.297) | | -0.0308 (0.171) |
| Constant | 0.582*** (0.201) | 0.429*** (0.107) | 0.683*** (0.204) | 0.654*** (0.213) | 0.377*** (0.121) | 0.370*** (0.124) |
| Observations | 1,284 | 1,284 | 340 | 340 | 944 | 944 |
| R-squared | 0.076 | 0.076 | 0.055 | 0.055 | 0.110 | 0.110 |

Notes: Regression results are based on Stage 1 fairness assessments only. 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 respondent.

Appendix P: Populated Pre-Analysis Plan

On September 21, 2020, we posted a pre-analysis plan on the AEA RCT Registry.⁵⁰ 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. 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.

⁵⁰ 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>.

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 | -0.0737 (0.0860) | | | | | |
| Taste-based × Employer | | -0.909*** (0.0679) | | | | |
| Statistical × Low-quality | | | -0.941*** (0.0745) | | | |
| Black discriminatee | | | | -0.347*** (0.0818) | -0.364*** (0.0911) | -0.279 (0.185) |
| Constant | -0.249*** (0.0722) | 0.132* (0.0784) | 0.221*** (0.0778) | -0.109 (0.0718) | -0.115 (0.0783) | -0.0890 (0.170) |
| 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 1. 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 3 (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.)⁵¹

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:

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

where BRO_i is respondent *i*’s assessment of Black peoples’ relative opportunities.⁵² *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 5 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.

⁵¹ Except in the small handful of cases where noted, this sample restriction has no effect on the results.

⁵² 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 | -0.00547 (0.174) |
| Other race respondent | -0.174 (0.220) |
| Male | 0.113 (0.127) |
| Age 35-44 | -0.0803 (0.149) |
| Age 45-54 | -0.262 (0.170) |
| Age 55 and over | -0.198 (0.233) |
| Democrat | -0.724*** (0.166) |
| Independent or other party | -0.522** (0.212) |
| Liberal | -0.728*** (0.194) |
| Moderate | -0.362** (0.178) |
| Four-year college | 0.314** (0.138) |
| Graduate School | 0.458** (0.190) |
| Constant | 3.703*** (0.210) |
| 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:

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

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 3 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 4 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 | -0.0892 | (0.0940) |
| Statistical × Low-quality info | -0.941*** | (0.0749) |
| Taste-based × Customer | -0.909*** | (0.0682) |
| Black discriminatee | 0.0646 | (0.250) |
| B. Respondent Characteristics: | | |
| Black respondent | 0.120 | (0.240) |
| Other race respondent | -0.265 | (0.234) |
| Male | 0.0522 | (0.147) |
| Age 35-44 | -0.0753 | (0.165) |
| Age 45-54 | -0.152 | (0.207) |
| Age55 and over | 0.0886 | (0.258) |
| Democrat | -0.447*** | (0.165) |
| Independent or other party | -0.615*** | (0.233) |
| Liberal | -0.364* | (0.200) |
| Moderate | -0.252 | (0.198) |
| Four-year college or university | 0.461*** | (0.161) |
| Graduate school | 0.832*** | (0.205) |
| C. Race Treatment Interactions with Respondent Characteristics: | | |
| Black Discriminatee × Black respondent | 0.0578 | (0.261) |
| Black Discriminatee × Other race respondent | 0.162 | (0.283) |
| Black Discriminatee × Male respondent | 0.200 | (0.161) |
| Black Discriminatee × Age35-44 | -0.132 | (0.195) |
| Black Discriminatee × Age45-54 | 0.0904 | (0.212) |
| Black Discriminatee × Age 55 and over | -0.576** | (0.269) |
| Black Discriminatee × Democrat | -0.211 | (0.186) |
| Black Discriminatee × Independent or other party | 0.0595 | (0.268) |
| Black Discriminatee × Liberal | -0.518** | (0.219) |
| Black Discriminatee × Moderate | -0.510** | (0.215) |
| Black Discriminatee × Four-year college | 0.0858 | (0.180) |
| Black Discriminatee × Graduate School | -0.231 | (0.227) |
| Constant | 0.534** | (0.255) |
| 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:

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

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.⁵³

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^2 s, 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.

⁵³ 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 | -0.102 (0.0937) | -0.0897 (0.0947) | |
| Statistical × Low-quality | -0.941*** (0.0746) | -0.941*** (0.0745) | |
| Taste × Employer | -0.909*** (0.0679) | -0.909*** (0.0679) | |
| White respondent | 0.344 (0.234) | | 0.341 (0.234) |
| Black respondent | 0.698** (0.330) | | 0.697** (0.330) |
| Black discriminatee × White Respondent | -0.367*** (0.0913) | | -0.364*** (0.0911) |
| Black discriminatee × Other race respondent | -0.122 (0.237) | | -0.120 (0.237) |
| Black discriminatee × Black respondent | -0.410 (0.281) | | -0.416 (0.280) |
| Constant | 0.0553 (0.228) | 0.221*** (0.0778) | -0.457** (0.220) |
| 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 | -0.0956 (0.0944) | -0.0207 (0.0984) | -0.160 (0.107) | 0.0127 (0.114) | 0.128 (0.198) | -0.121 (0.197) |
| Statistical × Low-quality | -0.941*** (0.0746) | -0.941*** (0.0861) | -1.020*** (0.0845) | -1.020*** (0.0975) | -0.660*** (0.158) | -0.660*** (0.182) |
| Taste × Employer | -0.909*** (0.0679) | -0.909*** (0.0784) | -0.887*** (0.0773) | -0.887*** (0.0893) | -0.986*** (0.142) | -0.986*** (0.165) |
| Black discriminatee | -0.348*** (0.0820) | -0.313*** (0.0726) | -0.367*** (0.0913) | -0.289*** (0.0825) | -0.278 (0.185) | -0.374** (0.154) |
| Constant | 0.401*** (0.0881) | -0.120* (0.0707) | 0.442*** (0.0964) | -0.135* (0.0816) | 0.258 (0.207) | -1.841*** (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).⁵⁴ 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.

In the paper, this “conditional utilitarianism” model (now called *Belief-Based Utilitarianism*, or *BBU*) is tested in Section 4.4. This model argues that respondents with utilitarian preferences (i.e. preferring to ‘help the underdog’) might base their fairness assessments on their personal *perceptions* of which racial group has more economic opportunities (BRO). Similarly to equation 5, our test regresses fairness perceptions on BRO, separately for discrimination against White versus Black job applicants. In contrast to equation 5, however, we use a continuous version of BRO (showing all 7 of its values) rather than a dichotomized version. The results are shown in Figure 6, separately for conservative and moderate/liberal respondents. The same key inconsistencies as in Table P2.4 lead us to reject this model for both political groups.

⁵⁴ Thus, BMO = 1 if the respondent chooses responses 4-7 on the 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 | -0.0837 (0.0919) | -0.152 (0.106) | 0.149 (0.181) |
| Statistical × Low-quality | -0.941*** (0.0746) | -1.020*** (0.0845) | -0.660*** (0.158) |
| Taste × Employer | -0.909*** (0.0679) | -0.887*** (0.0774) | -0.986*** (0.143) |
| BMO (δ^1) | 0.854*** (0.145) | 0.646*** (0.163) | 1.537*** (0.315) |
| BFO × Black discriminatee (δ^2) | -0.599*** (0.0964) | -0.665*** (0.107) | -0.383* (0.212) |
| BMO × Black discriminatee (δ^3) | 0.0421 (0.121) | 0.0980 (0.136) | -0.0888 (0.280) |
| Constant | 0.0856 (0.104) | 0.205* (0.115) | -0.321 (0.233) |
| 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 (0.0944) | -0.0207 (0.0984) | -0.0555 (0.140) | -0.126 (0.141) |
| Statistical × Low-quality | -0.941*** (0.0746) | -0.941*** (0.0861) | -0.970*** (0.0983) | -0.909*** (0.0965) |
| Taste-based × Employer | -0.909*** (0.0679) | -0.909*** (0.0784) | -0.875*** (0.0883) | -0.940*** (0.0865) |
| Black discriminatee | -0.348*** (0.0820) | -0.313*** (0.0726) | -0.505*** (0.129) | -0.192 (0.133) |
| Constant | 0.401*** (0.0881) | -0.120* (0.0707) | 0.492*** (0.114) | 0.306** (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 6 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. To that end, 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. Since the results are very similar, we did not pursue this issue further in the paper.

Table P3.1: A Look at “Thoughtful” Respondents

| | Full Sample (1) | Full Sample (2) | “Thoughtful” Sample (3) | “Thoughtful” Sample (4) |
|---------------------------|-----------------------|-----------------------|-------------------------------|-------------------------------|
| Taste-based | -0.0956 (0.0944) | -0.0207 (0.0984) | -0.119 (0.130) | -0.0335 (0.140) |
| Statistical × Low-quality | -0.941*** (0.0746) | -0.941*** (0.0861) | -0.764*** (0.0983) | -0.764*** (0.114) |
| Taste × Employer | -0.909*** (0.0679) | -0.909*** (0.0784) | -0.845*** (0.0970) | -0.845*** (0.112) |
| Black discriminatee | -0.348*** (0.0820) | -0.313*** (0.0726) | -0.449*** (0.114) | -0.247** (0.109) |
| Constant | 0.401*** (0.0881) | -0.120* (0.0707) | 0.690*** (0.121) | 0.166 (0.145) |
| 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 respondent.

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 7’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

- We decided not to include PAP1.6’s “actions versus identity” decomposition from the paper because it seemed of limited interest
- Due to a lack of statistical power, we did not pursue P3.1’s idea of using within-subject variation in responses to treatments to classify subjects into types.
- Table P3.1 replicates some of the key PAP results for a subset of ‘thoughtful’ respondents who took more than the median time to complete the survey. The results were very similar. For space reasons we decided not to conduct similar robustness tests on the paper’s main results.