When is Discrimination Unfair?

Pre-Analysis Plan

September 18, 2020

Our analysis will proceed in three stages, the first of which uses simple *t*-tests to establish some foundational facts about peoples' perceptions of discrimination. In stage two, we will explore the ability of some simple models of subjective fairness to account for these broad patterns, first by testing the implications of each 'pure' model, then by considering some hybrid models that might better summarize the data. Finally, in stage three we conduct a variety of robustness tests, heterogeneity analyses, and extensions.

In this document we describe our analysis plans for these three stages in decreasing levels of detail. Our plans for the later stages are less clearly formulated at this date, and will depend in part on our findings in the preceding stages.

STAGE ONE: ESTABLISHING THE MAIN FACTS

1.1 Is Taste-Based Discrimination Seen as Less Fair than Statistical Discrimination?

To test for an overall fairness difference between taste and statistical discrimination, we pool the White and Black treatments, the sub-types of Statistical and Taste-Based Discrimination (*low* versus *high quality* information, and *own* versus *customer tastes*), the two stages of the experiment, and respondents of all races. With an overall sample of 600 respondents, we will have 1200 evaluations of statistical discrimination and 1200 of taste-based discrimination.[1] We then conduct a *t*-test for the difference in means between the perceived fairness of these two types of discrimination, clustering the standard errors by respondent to adjust for within-subject error correlations.

Since the power of this test (and all the remaining ones) depends on the amount of within-subject error correlation, we bound the possible power levels by considering two extreme cases: zero correlation (in which case we have 2400 independent observations) and perfect correlation (in which case we have 600 independent observations). Minimum Detectable Differences (MDD's) under these two cases are presented in Table 1.

Based on our reading of the economics literature, we hypothesize that people, like economists, will tend to view taste-based discrimination as less fair than statistical discrimination.

1.2 How Do People Respond to Sub-types of Taste-Based and Statistical Discrimination?

We hypothesize that hiring discrimination based on the employer's own tastes (E) will be perceived as less fair than hiring discrimination that accommodates customers' tastes (C). While the discriminatory act (the hiring decision) in both cases is done by the employer, in the second case it is not motivated by

[1] We have funds for 714 respondents, but are using 600 as a conservative estimate of ultimate sample size to allow for unforeseen technical problems. With 600 respondents, 300 people will encounter two statistical scenarios each in stage 1 , and the same number will do so in stage 2.

the employer's own animus. We also hypothesize that statistical discrimination is viewed as less fair when it is based on low (L) versus high-quality (H) information about relative group productivity. Employers who hire based on low quality information may be relying on stereotypes rather than making the effort to actually measure relative qualifications.

To test for these differences, we again pool the White and Black treatments, the two stages of the experiment, and respondents of all races, but conduct separate tests for the statistical discrimination questions and the taste-based questions. With an overall sample of 600 respondents, we will now have 1200 evaluations of, say, statistical discrimination, with 600 of each sub-type. As always, *t*-tests for a difference between the sub-types will be clustered by respondent, and Table 1 calculates the bounds on the resulting MDDs.

1.3 Do People React Differently to Discrimination Against their Own Race versus Other Races?

To answer this question, we conduct separate analyses for two groups of respondents—those who self-identify as White and all other respondents (henceforth, White and Non-White).[2] We expect to have about 300 respondents in each group. If we pool all the types and sub-types of discrimination and the two stages of the experiment, we will have 600 evaluations of discrimination against Black people and 600 against White people for each of the two respondent-race groups.[3] Bounds on the implied MDD between observing a White or Black discriminatee on fairness assessments are again reported in Table 1.

For White respondents, we do not have a strong prior for the effect of the discriminatee's race because two plausible models of fairness described in Stage 2 have opposite predictions. Specifically, the *in-group bias model* predicts that White people will react more negatively to discrimination against White people than against Black people. A model with *utilitarian social preferences*, however, would predict the opposite, given that Black people have lower incomes and opportunities. For Non-White Respondents (and Black people specifically), both the in-group bias and utilitarian models predict they will react more negatively to discrimination against Black people than White people.

1.4 How Do Perceptions of Black and White Peoples' Relative Opportunities Vary with Race, Gender, Age, and Political Preferences?

Here we will pool all respondent races, all treatments, and both stages of the survey to obtain 2400 evaluations of discriminatory acts from 600 respondents. In this sample, we run the following regression:

$$BRO_i = \alpha + \theta^1 RR_i + \theta^2 RG_i + \theta^3 RA_i + \theta^4 RP_i + \varepsilon_{ij} \qquad (1)$$

where BRO_i is respondent *i*'s assessment of Black peoples' relative opportunities. *RR, RG, RA,* and *RP* represent (sets of) dummy variables for respondent race, gender, age, and political preferences, respectively. As always, ε_{ij} is clustered by respondent. We do not have strong priors for these effects,

[2] We do not expect to have a large enough sample of Black people to consider them separately, but will report these results in an Appendix.

[3] With an overall sample of 300 White respondents, 150 people will encounter two B scenarios each in stage 1, and the same number will do so in stage 2.

though we note that factors like in-group bias could generate motivated beliefs about relative opportunities.

1.5 How Does Racial Bias in Fairness Assessments vary with Race, Gender, Age, and Political Preferences?

Again in the full sample of 2400 fairness assessments, we'll now regress:

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

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.

1.6 What Matters More for the Perceived Fairness of Discrimination: Actions or Identity?

As a final descriptive exercise, we again pool all respondent races, all treatments, and both stages of the survey to obtain 2400 evaluations of discriminatory acts from 600 respondents. In this sample, we run the following regression:

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

Again, 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.[4] We will estimate equation (3) three different ways: as is, and using only the procedural or identity covariates alone. Together, the R^2s of these regressions will tell us which set of factors explains most of the variation in perceived fairness of discrimination.

[4] As already noted, in most of our analysis we will 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 1: Minimum Detectable Differences for Selected Tests of the Main Facts

Test:	MDD (worst case)	MDD (best case)
1. Taste versus Statistical Discrimination	0.2286	0.1143
2. Difference between Sub-Types of Discrimination	0.3233	0.1617
3. Effect of a Black Discriminatee, for fixed Respondent Race	0.3233	0.1617

Note: All calculations are based on a total of 600 respondents. MDDs are measured in standard deviations and are calculated as $MDD = \left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)+\Phi^{-1}(\beta)\right]\sqrt{\frac{2}{m}}$ where m denotes the number of observations in a comparison group and Φ is the standard normal cdf. This equation assumes that the comparison groups have the same variances and are balanced. For our calculations, we let $\alpha = 0.05$ and $\beta = 0.8$. The "worst case" assumes that errors are perfectly correlated within subjects, while the best case assumes that they are independent within subjects.

STAGE TWO: MODELS OF FAIRNESS—MAKING SENSE OF RESPONDENTS' FAIRNESS PERCEPTIONS

In this Stage of the project, we begin by testing three simple, polar-case models of how respondents might judge the fairness of discriminatory actions. All three tests are applied to 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. Because the predictions of models 1 and 2 do not depend on the respondent's race, our main tests of those models will pool respondents of all races.[5] Model 3 (in-group bias), however, predicts that the coefficients of equation (4) should vary with the race of the respondent. Here, our main tests will be separate regressions for White versus Non-White respondents, as described in Stage 1.3 above.

After testing these three 'pure' models, we will assess the ability of a variety of hybrid models to better account for the patterns in the data. Since we are still thinking about the best way(s) to do this, the current document only provides examples of the types of models we might consider, depending in part on earlier results.

2.1 The Utilitarian Social Preferences Model

A survey respondent who bases her fairness assessments purely on utilitarian social preferences cares about *outcomes* (in this case, who was denied the job) and not on *process* (the actions and motivations that lead to the outcome). Further, a utilitarian will prefer outcomes that redistribute income or opportunities from people with higher incomes to those with lower incomes.

Since Black people have demonstrably lower incomes than White people in the United States, a pure Utilitarian Social Preferences model predicts $\beta = 0$ and $\delta < 0$ (i.e. discrimination against Black people is less fair than discrimination against White people). This should be true for survey respondents of all races if those respondents have utilitarian social preferences.

2.2 The Rules-Based Fairness Model

A survey respondent who bases her fairness assessments purely on rules that specify which actions are acceptable or not, does not care about outcomes or about the identities of the people involved, but only on *process* (the actions taken by the employer in our scenarios). For example, deciding not to hire someone because you prefer not to interact with them would be considered equally wrong, regardless of the races of the people involved.

Thus, the pure Rules-Based Fairness model predicts $\beta \neq 0$ and $\delta = 0$: the races of the discriminator and discriminatee should be irrelevant. Notice that the Rules-Based Fairness model makes no predictions about the overall acceptability of discriminatory acts: A person who adopts these principles could believe, for example, that statistical discrimination is always wrong, or that it is always acceptable. The

[5] We will, however, also estimate equation (4) separately by respondent race to test the prediction of the utilitarian model that respondents *of every racial group* should be less tolerant of discrimination against Black than White people, and to see whether other norms for ethical behavior differ across racial groups.

important thing is that the acceptability of a given act does not depend on the identities of the people involved.

An additional prediction of the Rules-Based Fairness model applies if we extend equation (4) to include interactions between actions (A) and identities (B): since the same rules (about which types of discrimination are more objectionable than others) should apply to all people, these interaction effects should be zero as well. Again, all these predictions should hold for survey respondents of all races, as long as respondents follow the rules-based fairness model.

2.3 The In-Group Bias Model

A survey respondent whose fairness assessments are driven purely by in-group bias is less accepting of actions that hurt a member of their in-group than actions that hurt an out-group member. Since only outcomes (the amount and distribution of harm) matter here, we expect $\beta = 0$.[6] In contrast to the two previous models, however, our predictions for the sign of δ now depend on the race of the survey respondent. Specifically, we should see $\delta > 0$ for White respondents (i.e. White people will rate discrimination against Black people as more fair than discrimination against White people), and $\delta < 0$ for Non-White respondents.[7]

2.4 A Hybrid Model: Conditional Utilitarianism

We plan to explore the ability of various combinations of three preceding polar-case models of subjective fairness to account for the patterns in our data. This may involve writing down a more formal model in which subjects rely on all three of these principles to different degrees. At the time of writing, we do not yet have a clear idea the most practical way to do this. As an example of the kinds of models we might consider, we describe a *conditional utilitarianism* model and its testable implications here. This model combines utilitarianism, in-group bias, and subjective beliefs about relative opportunities.

In a conditional utilitarianism model, different respondents may have different perceptions about whether Black or White people have more economic opportunities. These beliefs may be correct or incorrect, and may or may not be motivated by in-group bias. Either way, the conditional utilitarian model posits that respondents' fairness assessments are consistent with their stated beliefs about relative opportunities.

We illustrate our test of the conditional utilitarian model for the case of White respondents, and start by tentatively dividing White people 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).[8] We then expand equation (4) to include interactions between the Black treatment (*B*, where the discriminatee is Black) and BMO, as follows:

---

[6] Notice that the amount of harm inflicted by the discriminatory act is the same for all four acts we consider: the discriminatee did not get the job.

[7] The in-group bias hypothesis actually predicts $\delta < 0$ for Black respondents and $\delta = 0$ for "Other" racial groups, if those groups view neither White nor Black people as their in-group. To the extent that our sample size permits, we will explore conducting this test using only Black respondents as well.

[8] 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

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

In equation (2), 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 (White people in this case) whom they *believe* have fewer opportunities—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 ($\delta^3 > 0$).

A useful feature of the conditional utilitarian model is that in some cases, it allows us to distinguish 'pure' in-group bias from in-group bias that is motivated or supported by inaccurate beliefs. To see this, suppose that our baseline regression, equation (4) showed that White people on average, were more accepting of discrimination against Black people than White people ($\delta < 0$), suggesting the presence of in-group bias among White people. By estimating equation (5) we can distinguish whether this preference occurs *despite* a belief that Black people have fewer opportunities (the case of 'pure' in-group bias), or whether it is supported by inaccurate or motivated beliefs that Black people, in fact have more opportunities.

In sum, equation (5) can tell us whether respondents' patterns of fairness assessments can be explained by utilitarian social preferences that are consistent with respondents stated beliefs about whether White or Black people have more economic opportunities. If so, respondents' choices can be rationalized by their beliefs plus a desire to help the "underdog" as they perceive it to be. If not, 'pure' in-group bias may also play a role.

2.5 Interactions between Distributional Considerations and Concerns for Procedural Fairness

As a final illustration of how our experimental data can be used to assess some more nuanced determinants of perceived fairness, we could leverage 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. To do so, we would 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 between-subject estimates of δ from stage 1 of the survey only—this would suggest that subjects also care about consistency.

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.

---

opportunities because we expect the latter group to be considerably smaller in size. We may explore other cut-offs as well if the median answer to this question is far from "equal opportunities".

We stress that models 2.4 and 2.5 are only examples of the hybrid and extended models of fairness we will consider in Stage Two of our analysis. Indeed, models 2.4 and 2.5 may not even be relevant, depending on our Stage One results. For example, model 2.4 (conditional utilitarianism) will be largely irrelevant if we discover that virtually all respondents of all races have similar beliefs about Black people' relative opportunities. In sum, after testing the three polar case models (2.1-2.3), our investigation will be more inductive in nature, aiming to identify one or more simple models that are consistent with the main patterns we detect. This investigation will be guided, in part, by our open-ended survey question, where people are asked to explain the reasons for one of their fairness assessments.

STAGE THREE: ROBUSTNESS AND HETEROGENEITY

3.1. Heterogeneity

In addition to considering hybrid models of subjective fairness, we plan to make some efforts to assess whether and how different populations of subjects act according to different models of fairness. We are considering two approaches to this question, the first of which uses within-subject estimates to classify individual respondents. While we expect to have only very limited statistical power for this exercise, we may be able to make some headway by exploiting the fact that the race of the discriminatee changes for two thirds of our respondents between survey stages 1 and 2.[9] Thus, we could (crudely) define, say, in-group biased White people as those who reacted more negatively to discrimination against White people than Black people, then compare the demographics of this sub-group of White people to other White people.

A second approach to heterogeneity analysis would be to divide the respondents into large sub-samples, and replicate equation (4) (which relies on both within- and between-subject variation) on these sub-samples. The most obvious sample divisions would seem to be:

- 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[10]

[9] For half of those switchers, the discriminatory actions change too, so that would need to be adjusted for.
[10] 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.

3.2 Robustness

While our main analysis will standardize the subjective fairness scores in a simple way (relative to the mean and standard deviation of all fairness assessments combined), we may explore alternative standardizations.

Alternative cut-offs for defining discrete categories, such as the cut-off between BMO and BFO, will be explored.

While most of our analysis will use two respondent race categories (White versus Non-White) we may explore more detailed groupings as well (though our power is likely to be limited).

We will explore if the results change when we restrict attention to more 'thoughtful' subjects who took more than a minimum number of seconds to click on their fairness assessments.