

UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Essays in Applied Labor Economics

Permalink

<https://escholarship.org/uc/item/0zw1t8wc>

Author

Osaki, Trevor Tamotsu

Publication Date

2023

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

Essays in Applied Labor Economics

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Economics

by

Trevor Tamotsu Osaki

Committee in charge:

Professor Kelly Bedard, Chair
Professor Peter Kuhn
Professor Gonzalo Vazquez-Bare

June 2023

The Dissertation of Trevor Tamotsu Osaki is approved.

Professor Peter Kuhn

Professor Gonzalo Vazquez-Bare

Professor Kelly Bedard, Committee Chair

June 2023

Essays in Applied Labor Economics

Copyright © 2023

by

Trevor Tamotsu Osaki

To my mom and dad, and my twin sister, Jennifer.

Acknowledgements

I want to thank my Ph.D. committee members, Kelly Bedard, Peter Kuhn, and Gonzalo Vazquez-Bare, for their guidance and support throughout the development of this dissertation. Kelly always reassured me of my ability to push forward while writing my dissertation and make each chapter a reality. Peter's mentorship helped improve my technical writing skills by a significant amount since I arrived at UCSB. And Gonzalo always provided me with insightful writing suggestions and tips for applying empirical methods. I am grateful to have these individuals as my committee members, and I could not have made it this far without them.

I would also like to thank my mom and dad and my twin sister, Jennifer. They constantly cheered me on and never doubted my ability to reach each major milestone, from passing the first-year preliminary exams to successfully navigating through the job market. Their love and support motivated me to trek onward throughout my graduate career. Therefore, I greatly owe my triumphs to them.

Finally, I thank my cohort mates, colleagues, and friends. Because of them, my experience at UCSB could not have been any more enriching, fun, and memorable.

Curriculum Vitæ

Trevor Tamotsu Osaki

Education

- 2023 Ph.D. in Economics (Expected), University of California, Santa Barbara.
- 2018 M.A. in Economics, University of California, Santa Barbara.
- 2017 B.A. in Economics, Occidental College

Research

The Effects of Test-Optional Admissions on Underrrepresented Minority Enrollment and Graduation

When Is Discrimination Unfair? (with Peter Kuhn)

Do Perkins Loans Come With Perks?

Teaching Experience

- 2017 - 2023 Teaching Assistant, University of California, Santa Barbara (UCSB) Department of Economics
- 2016 - 2017 Facilitator, Academic Mastery Program, Occidental College

Service

- 2020 - present Advisor, Pathway to Success, Cerritos College
- 2021 Advisor, Senior Honors Thesis (ECON 196), UCSB Department of Economics

Presentations

- 2022 Consortium of Applied Research in Economics Seminar (UCSB)
- 2022 Applied Microeconomics Lunch (UCSB)
- 2022 David P. Gardner Seminar (University of California, Berkeley)
- 2022 Trans-Pacific Labor Workshop
- 2021 Applied Microeconomics Lunch (UCSB)

Awards & Honors

2022	Outstanding Undergraduate Course TA (Winter Quarter), UCSB Department of Economics
2021	David P. Gardner Fellowship, Center of Studies in Higher Education, University of California, Berkeley
2019	Research Fellowship, UCSB Department of Economics
2017	Graduate Recruitment Fellowship, UCSB Department of Economics
2017	Phi Beta Kappa, Delta Chapter of California, Occidental College
2016	Ford Research Mentor's Endowment Fellowship
2010	Eagle Scout, Boy Scouts of America

Skills

Advanced: Stata, R, \LaTeX , and Microsoft Office (Word, Excel, & Powerpoint)

Basic: Matlab

Abstract

Essays in Applied Labor Economics

by

Trevor Tamotsu Osaki

This dissertation includes three chapters in applied labor economics. The first chapter considers the growing trends of colleges and universities making the submission of SAT and ACT scores optional for undergraduate admissions to bolster racial diversity. It uses a panel of liberal arts colleges from IPEDS and applies a two-way fixed effects approach to determine whether this test-optional policy is effective at achieving this goal. It also estimates the impact of the policy on the graduation rates for underrepresented minority (URM) students. This chapter finds that the policy bolsters freshman URM enrollment among test-optional institutions throughout the sample as a whole, regardless of admissions selectivity and early versus late treatment timing. The effects of this policy on the URM 4-year and 6-year graduation rates are heterogeneous across colleges by their selectivity in admissions. While the most selective colleges in the panel experience no change in URM graduation rates, less-selective colleges experience declines in these graduation rates.

The second chapter, joint with Peter Kuhn, uses a vignette-based survey experiment on Amazon's Mechanical Turk to 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, survey subjects react in very similar ways to the employer's *motivations* for the action, such as the quality of information on which statistical discrimination is based. Compared to conservatives, moderates and liberals are much less accepting of

discriminatory actions and consider the discriminatee’s race when making their fairness assessments. This chapter also describes four pre-registered models of fairness – utilitarianism, race-blind rules (RBRs), racial in-group bias, and belief-based utilitarianism (BBU) – and shows that the latter two are inconsistent with major aggregate patterns in the survey data. Instead, it argues that a two-group framework, in which one group (mostly self-described conservatives) values employers’ decision rights and the remaining respondents value utilitarian concerns, explains the main findings well. In this model, both groups also value applying a consistent set of fairness rules in a race-blind manner.

The third chapter estimates the effects of the Federal Perkins Loans Program on students’ non-academic work status and labor supply, as well as their choice of STEM major and post-graduation occupation via matching. Perkins loans were unique because they converted into grants if a student recipient entered a designated career path. Therefore, this chapter leverages this conditional loan cancellation feature to determine whether recipients enjoyed the benefits often associated with grant programs (e.g., reduced incentive to work) and motivated them to enter loan-forgiving occupations. It finds that Perkins loans reduce students’ non-academic work activity and grades. The effect size for the latter outcome, however, is marginally small. And this chapter finds that the program had little impact on their likelihood of pursuing a STEM major or loan-forgiving occupation.

Permissions and Attributions

Chapter 2 and Appendix B of this dissertation is the result of collaboration with Peter Kuhn.

Contents

Curriculum Vitae	vi
Abstract	viii
1 Effects of Test-Optional Admissions on URM Enrollment and Graduation	1
1.1 Introduction	1
1.2 Background and Potential Mechanisms	5
1.3 Data and Sample	8
1.4 Empirical Strategy	11
1.5 Results	14
1.6 Discussion	20
1.7 Figures	24
1.8 Tables	28
2 When Is Discrimination Unfair?	32
2.1 Introduction	32
2.2 Design and Implementation	38
2.3 Some Facts	45
2.4 Assessing Four Models of Fairness	50
2.5 Reconciling Conflicting Fairness Criteria: Utilitarianism versus RBRs	61
2.6 Robustness	66
2.7 Discussion	71
2.8 Figures	76
3 Do Perkins Loans Come With Perks?	83
3.1 Introduction	83
3.2 Background of the Perkins Loan Program	86
3.3 Data	88
3.4 Empirical Strategies	91
3.5 Results	97

3.6	Discussion	99
3.7	Figures	102
3.8	Tables	103
A	Additional Material for Chapter 1	109
A.1	Sample Details	110
A.2	Results without Control Variables	113
A.3	Propensity-Trimmed Sample	117
A.4	Heterogeneity Robust Estimation	126
A.5	Effects of the Policy on Non-first-time students	131
A.6	Enrollment of First-time Non-URM Students	134
A.7	Inclusion of HBCU Institutions	137
A.8	Non-URM Graduation Rates	140
A.9	Analysis at the State Level	143
A.10	Miscellaneous Figures and Tables	145
B	Additional Material for Chapter 2	147
B.1	Survey Design	147
B.2	Representativeness	157
B.3	Order Effects	164
B.4	Exploring the Effects of Education on Fairness Ratings	178
B.5	Robustness Tests for Sections 3 and 4	182
B.6	Analysis of Open-Text Responses	188
B.7	Experimenter Demand Effects do not Explain the Race Treatment Order Effects	194
B.8	Estimating α	198
B.9	Replicating the Main Figures with ACS Weights	208
B.10	Replicating the Main Figures with GSS Weights	215
B.11	Replicating the Main Figures with Standardized Fairness Measures	222
B.12	Replicating the Main Figures for ‘Thoughtful’ Subjects Only	228
P	Populated Pre-Analysis Plan	235
C	Additional Material for Chapter 3	261
C.1	More on B & B	261
C.2	Predicting Perkins Take-up	265

Chapter 1

Effects of Test-Optional Admissions on URM Enrollment and Graduation

1.1 Introduction

Racial and socioeconomic gaps in higher education attainment have been prevalent throughout the past several decades (Page and Scott-Clayton, 2016). The economic literature that addresses this issue includes studies that assess the impact of various policies intended to level the playing field, such as federal financial aid programs. One such policy, test-optional admissions, has received much less attention within this literature.

Many selective 4-year colleges and universities in the U.S. rely on SAT and ACT scores to assess students for undergraduate admission. However, a growing number of schools are making the submission of these test scores optional.¹ Although some of these schools reason that standardized test scores are unreliable indicators of college preparedness, many of them also suggest their new admissions policy can help promote racial diversity among their enrolled student bodies. Some test-optional schools cite anecdotal

¹Specifically, the 2019 COVID-19 pandemic led the vast majority of colleges and universities to make the submission of standardized test scores optional due to the increased difficulty of facilitating proctored exams. Some of these schools effectively eliminated this requirement for admissions consideration. As discussed in Section 1.4, however, this study's scope does not encompass the pandemic.

evidence of their policy’s success in reaching this goal. For example, Providence College reports that its test-optional policy led to a 19 percent increase in underrepresented minority (URM) enrollees as well as a 56 percent increase in Pell-eligible enrollees (Epstein 2009).² As the number of test-optional schools grows, especially throughout the COVID-19 pandemic, the understanding of this policy’s effects on diversity and graduation outcomes is becoming increasingly important.

This study uses a panel of private liberal arts colleges and applies two-way fixed effects to determine whether this test-optional policy successfully bolstered racial diversity at these institutions.³ This study also examines the impact of the policy on the graduation rates for URM students, i.e., the share of URM students from a cohort who graduate within a specific duration of time. Liberal arts colleges are of interest since they primarily award 4-year undergraduate degrees (i.e., at least 50% of the degrees they offer to students), and a larger share of them dropped the test requirement within the panel’s time frame compared to other types of institutions.⁴

Proponents of the test-optional policy argue that standardized tests serve as admission barriers for prospective underrepresented students, resulting in racial and socioeconomic gaps in test scores. For example, *The Brookings Institute* reports that only 5% of Hispanic/Latino and 2% of Black students, compared to 60% Asian and 33% White students, are among the top-scoring SAT test-takers.⁵ These gaps stem from a few channels. First, there are inequalities in access to resources that can help applicants within college ad-

²URM students refer to those from Black, Hispanic/Latino, or American Indigenous racial groups. These groups are noted to be underrepresented at selective colleges and universities.

³The test-optional policy is distinct from other similar policies, such as test-blind admissions (i.e., colleges do not consider test scores at all). Although the effects of this admission regime may be worth investigating, this study excludes test-blind institutions to isolate the effects of the test-optional policy. The data source used by this study contains information on only one such institution: Hampshire College.

⁴Liberal arts colleges compete with other types of 4-year institutions, e.g., private and public R1 and R2 universities, for undergraduate students seeking Baccalaureate degrees.

⁵Richard V. Reeves and Dimitrios Halikias, *Race gaps in SAT scores highlight inequality and hinder upward mobility* (Washington D.C.: The Brookings Institute, 2017).

missions. Students from affluent backgrounds tend to have greater access to resources that can help them improve their scores on any standardized test via hiring a private tutor or registering for several administrations of the same exam (Vigdor and Clotfelter, 2003). Consequently, the SAT/ACT serves as a sorting mechanism that favors wealthy students and reinforces their disproportionate presence at the nation's most selective institutions (Anlon, 2009). These inequalities also lead to gaps in test scores across racial groups. Second, standardized college entrance exams are culturally biased (Freedle, 2003; Santelices and Wilson, 2010).⁶

There is some debate regarding the consequences of test-optional admissions. The most prominent point of contention toward this policy is the argument it could cause a decline in the academic preparedness of enrolling students (Epstein 2009). This corresponds to a phenomenon known as the *mismatch effect*, which posits that diversity-enhancing policies, such as race-blind admissions, can sort URM students into selective colleges and universities that would have otherwise rejected them. It hinges upon the assumption that some of these students tend to have lower academic credentials than their non-URM counterparts (Rothstein and Yoon, 2008). Consequently, upon enrollment, these students face peers that are much more academically prepared and will end up academically underperforming at their school of enrollment. Therefore, an analysis of the policy's effects on the graduation outcomes of URM students may, to some extent, shed light on whether the mismatch is a concern within a test-optional admissions regime.⁷

This study finds that test-optional admissions increase the volume of enrolling freshman students by 12.5%. Similarly, the policy increases the fraction of enrolling first-year

⁶Santelices and Wilson (2010) define *cultural bias* as the case in which cultural subgroups may interpret test items differently, especially in the case of *verbal* questions.

⁷The literature on mismatch has mixed results, with many of the utilized methodologies receiving heavy scrutiny. Arcidiacono et al. (2016) find robust evidence of school mismatch among STEM majors with the use of micro-data and counterfactual simulation. However, the appropriate methods needed to precisely identify mismatch exceed the scope of this study.

students from a URM background by over 23% from the pre-treatment mean. These positive URM enrollment effects are evident across the entire sample, regardless of institutions' selectivity in admissions and whether they dropped the test requirement earlier versus later in the panel's time frame. It also finds that the policy has a negligible impact on the 4-year and 6-year URM graduation rates among highly selective colleges. On the other hand, relatively less-selective colleges experience a roughly 10% decline in these graduation rates. This result raises the possibility of college mismatch at these less-selective institutions.

A few studies use panel data sets of colleges & universities to estimate the impact of the policy on the overall share of students from a URM background on campus, and they find that it has little effect on this outcome (Belasco et al., 2015; Saboe and Terrizzi, 2019). Another recent study, Bennett (2022), uses a propensity-matched sample of colleges and universities from 2005 to 2015. It finds that the policy led to a minor increase in the share of URM students and the volume of applicants.

This study contributes to the literature in a few ways. First, it utilizes a sample incorporating a much larger share of test-optional schools (i.e., treated units). Thus, it backs out a relatively precise policy effect estimate. Second, unlike a few papers from the literature, this study estimates the policy's impact on the outcomes for *first-year* URM students (e.g., the fraction of first-year students from a URM background). These variables are likely informative since they are closely connected to the college admissions process and may be very responsive to the policy. Third, this study assesses the policy's impact on the graduation outcomes of URM students to determine whether students benefitting from the policy are more or less likely to persist beyond the first year of college. To the best of my knowledge, this is the first paper in the economic literature to assess this policy's impact on URM students' graduation outcomes. Finally, this study exploits heterogeneity in the policy's effects across colleges' degree of admissions

selectivity.

This paper proceeds as follows. Section 1.2 discusses some background of the test-optional policy as well as the mechanisms that can explain how it could affect diversity outcomes of interest. Section 1.3 describes the sources of the data used in the primary analyses. Section 1.4 discusses the empirical strategy used to identify the effects of the policy. It also discusses the dynamic treatment effects of the policy as a test for identifying assumptions. Section 1.5 discusses the estimated effects of the test-optional policy. Finally, Section 1.6 concludes.

1.2 Background and Potential Mechanisms

The growing collection of schools that drop their standardized test requirement for admissions is informally known as the *test-optional movement*. This movement has primarily been driven by discourse that questions the usefulness of the SAT and ACT as proxies of student ability and cites their obstructive nature towards access to the most selective institutions. There indeed exists literature that substantiates these concerns about standardized tests. For example, Rothstein (2004) argues that the SAT's predictive power is lower than what the previous literature has suggested and recommends that it be assigned less importance within the admissions process. In addition, studies such as Blau, Moller, and Jones (2004), find that schools' reliance on standardized test scores can deter otherwise high-ability Black students from applying.

These concerns led Bowdoin Colleges and Bates College to drop their test score requirements in 1969 and 1984, respectively. Bates College, in particular, adopted it out of concern that its average SAT scores deterred strong students from applying (Epstein, 2009). It reported that this policy raised the number of applications it received and, at the same time, did not diminish the quality of its enrolling student cohorts. But it also

found that the number of Black and Hispanic/Latino applicants increased, with almost half choosing not to submit their test scores. The success of early adopters, such as Bates College, led many other schools to drop their test requirements. Anecdotal evidence from test-optional colleges has caught the attention of other schools and has led them to adopt the policy. For example, the University of Chicago’s vice president James Nondorf stated upon the university’s recent adoption of its test-optional policy, “[the university’s initiative] will further remove barriers to selective schools for students from underrepresented communities, including Pell applicants.”⁸ Other universities, such as Virginia Commonwealth University and George Mason University, have adopted the policy with the same intention.⁹

As shown in Figure A10.1, when a student navigates through the application for a test-optional college, they will be prompted to indicate whether they choose to submit their SAT or ACT score for admissions consideration. If a student opts to submit their test score, then it will be reviewed as a part of their application. Otherwise, if they do not, the college will heighten the importance of the remaining admissions criteria, such as high school GPA, letters of recommendation, or extracurricular activities. The fine print behind these colleges’ admission processes is unobservable. However, the way each institution approaches test-optional admissions may vary. For example, as suggested by Figure A10.2, some colleges indicate that they provide equitable admission consideration to students regardless of whether they choose to submit their test scores or not. If an applicant withholds their test score, the college would infer their academic ability using remaining admissions criterion without penalizing them in any form. Other colleges indicate to applicants that with the absence of test scores, they will place greater weight on specific criteria when making an admission decision. For example, as shown in Figure

⁸James G. Nondorf, “The University of Chicago, on Diversity,” *The New York Times*, July 13, 2018.

⁹Joey Matthews, “VCU to Drop SAT Requirement,” *Richmond Free Press*, January 1, 2015.

A10.3, Denison University places additional weight on applicants' coursework rigor if they choose not to submit their test scores.

From an ex-ante perspective, the effects of the test-optional policy on the admission of URM students are ambiguous. As a first possibility, the policy could positively affect URM admission and enrollment. When colleges provide equitable admission consideration between submitters and non-submitters, they can curtail the issues of gaps in SAT and ACT scores. Thus, if many URM applicants are non-submitters, they can benefit from the policy and, therefore, face better probabilities of admission. Furthermore, the policy's positive impact on URM student enrollment could take the form of a *warming effect*. In this case, URM students value campuses that foster racial and ethnic diversity and gain the most utility from applying and enrolling at campuses in the presence of peers with similar backgrounds (Card and Krueger, 2005). Test-optional institutions may be signaling to prospective URM students that they are attempting to foster campus diversity by adopting the policy, so these students may be likely to apply and enroll there. Thus, this effect would bolster the volume and share of these students enrolling at these schools.

As a second possibility, the policy may have a negligible impact on URM admission and enrollment. As previously discussed, test-optional admissions may place greater weight on specific admission criteria, such as the availability of college preparatory coursework (e.g., AP/IB programs) or extracurricular activities. However, some school-specific criteria, such as the availability of college preparatory coursework and extracurricular activities, are noted to be unequally distributed across demographic groups (Espenshade & Radford, 2009; Iatarola, Conger, & Long, 2011; Klugman, 2013; Perna et al., 2013). Thus, these test-optional colleges could be replacing one inequitable set of criteria with another within their admissions processes. Consequently, not all URM applicants may be able to benefit from test-optional admissions.

As a final possibility, this policy could harm URM admission through mechanisms akin to signaling (Spence, 1973) and models of school admission (Avery & Levin, 2010; Pop-Echeles & Urquiola, 2013).¹⁰ A college may desire to admit applicants with unobservable ability levels above some cutoff, but it can only use noisy proxies such as grades and test scores to infer applicants' abilities. Under a test-optional regime, the college will only be able to use grades to infer a student's ability if they choose to withhold their test score. However, the strongest applicants with high test scores may submit them anyway to signal to the college that they meet the desirable ability level. But consequently, the college may wrongly infer students withholding test scores have some ability level below the cutoff, so they would face lower probabilities of admission. Thus, this policy would harm the admission and enrollment of URM students if they tend to withhold their test scores.¹¹

Indeed, some of the mechanisms may be working in tandem with each other. Therefore, the sign of the estimated effects of the policy could roughly indicate which of these are predominant within the admissions and enrollment processes at test-optional schools.

1.3 Data and Sample

This study uses institution-specific panel data from the National Center of Education Statistics' Institutional Postsecondary Education Data System (IPEDS). The panel spans

¹⁰These two studies consider the implications for applicants when they choose to reveal a piece of information (e.g., preference-based information) that may affect their placement relative to some admissions cutoff. For example, Avery and Levin (2010) consider early admissions within the context of college admissions. And Pop-Echeles & Urquiola (2013) consider students' choice of academic track within the Romanian secondary school system.

¹¹This framework suggests other possible outcomes, or "equilibria." In particular, students with sufficiently high grades but low test scores could reveal them anyway to signal to college admissions that they possess the desired ability level. Hence, college admissions may infer that any student submitting their test scores is "high ability," whereas non-submitters would be inferred as "low ability."

the academic years 2001-2002 through 2019-2020.¹² The data set includes the number of first-time URM students enrolling at each institution, the fraction of first-time students from a URM background, and the 4-year and 6-year URM graduation rates.¹³ The data set also includes several control variables: logged full-time enrollment (FTE), logged tuition and fees, logged education and related (E & R) expenditures per FTE, logged institutional student grant aid per FTE, and a set of binary variables that reflects the extent college-preparatory classes are considered in admissions (e.g., college prep classes are recommended or required).¹⁴

The complete list of test-optional institutions is obtained from *FairTest*, an organization that addresses fairness and accuracy issues within U.S. student test-taking. Although this list is convenient for identifying the institutions belonging to the test-optional (treatment) group, it does not provide the dates they adopted the policy. However, the exact adoption dates are obtained from IPEDS.¹⁵ These dates are based on a yearly categorical variable that rates the extent to which test scores are considered in admissions.¹⁶

The sample includes a total of 149 liberal arts colleges, of which about 45% adopted the policy sometime between 2002-2003 and 2019-2020, respectively. Since test-optional

¹²The choice of the initial period, 2001-2002, reflects data availability from IPEDS for some key variables. Although IPEDS includes periods beyond 2019-2020, the data set for this study does span further since the vast majority of schools dropped the test requirement during the 2020-2019 admission cycle due to the COVID-19 pandemic. This would essentially nullify the identification strategy discussed in Section 1.4 (i.e., almost all schools would be treated).

¹³*First-time students*, as defined by IPEDS, are students that have no prior postsecondary experience attending any institution for the first time at the undergraduate level. Therefore, none of these students transferred from a 2-year institution. Similarly, the outcomes on graduation solely reflect individuals that enrolled as first-time students.

¹⁴All monetary control variables are logged and adjusted for inflation in 2019 USD. Furthermore, only a few colleges seem to have adjusted their consideration of college preparatory courses simultaneous with dropping the test-requirement. However, the sample size is relatively small and, therefore, sensitive to outliers. Thus, these binary variables are included in the specification, although their inclusion changes the point estimates by a very small amount.

¹⁵The set of schools suggested by IPEDS to be test-optional is consistent with that of FairTest.

¹⁶This variable takes on a value of “1” if test scores are required, “2” if they are recommended, “3” if they are neither required nor recommended, or “5” if they are considered but not required. Any institution whose variable takes on a value of 2, 3, or 5 is considered test-optional.

institutions dropped the test requirement in various years, policy adoption is “staggered.” Figure 1.1 contains the distribution for the years in which institutions from the sample adopted the policy. The inclusion of liberal arts colleges within the analyses ensures the comparability of all units across the treatment and control groups since a significant fraction of test-optional colleges through 2019-2020 belong to this category of institutions.¹⁷ All of these institutions are either considered to be “selective,” “more selective,” or “most selective” by the U.S. News and World Report (USNWR). None of them are Historically Black Colleges and Universities (HBCUs).¹⁸ Although IPEDS contains data on a larger number of institutions, many were discarded due to a substantial number of missing observations for some key variables. A large number of these excluded institutions were established after 2001-2002. Furthermore, all test-optional institutions (e.g., Bates College) that adopted their policy before 2001-2002 were dropped from the sample.¹⁹ As a reference, Table A1.1 from Appendix A.1 contains the list of test-optional institutions in this sample and their year of policy adoption.

The summary statistics for the outcome and control variables are displayed in Table 1.1. These statistics correspond to the 2001-2002 (pre-treatment) observations of the displayed variables. Columns (1) and (2) contain the statistics for the test-optional and test-requiring institutions, respectively. Column (3) includes p -values for the difference

¹⁷There are other types of institutions that adopted test-optional admissions, such as doctoral universities. However, these institutions differ from liberal arts colleges in many ways, such as endowment, enrollment level, etc. Therefore, these institutions may not serve as proper control units with the analyses. Furthermore, most of these types of institutions dropped the test requirement within the last several periods leading up to 2019-2020. On the other hand, the timing of treatment among liberal arts colleges has more variation within the time frame of the panel.

¹⁸A preliminary sample of institutions from IPEDS contains three HBCUs, none of which are test-optional. Appendix A.7 reproduces much of the analyses using a sample that includes these institutions. It shows that the inclusion of HBCUs do not significantly change the results from the body of this paper.

¹⁹IPEDS has data on a total of seven colleges thought would be considered as “always-treated” units. As discussed in the next section, this study estimates the effects of the policy using a TWFE approach with staggered treatment. The estimator for these effects is a weighted average of average treatment effects on the treated (ATT) (Goodman-Bacon, 2020). When always-treated units are included within such regression, they are treated as control units and are given disproportionately greater weight. If their treatment effects change over time, they bias TWFE estimator from a meaningful parameter.

in means between the test-optional and requiring institutions for all variables. These differences, except for a few control variables, are statistically indistinguishable from zero.

1.4 Empirical Strategy

1.4.1 Specifications

This study utilizes a two-way fixed effects (TWFE) approach. The institution’s choice to adopt the test-optional policy serves as the “treatment.” As discussed in Section 1.3, treatment is staggered across academic years. Given institution i in academic year $t \in \{2001-2002, \dots, 2019-2020\}$, the primary specification is represented by the following equation:

$$y_{it} = \beta P_{it} + \mathbf{X}'_{it}\gamma + \alpha_i + \lambda_t + \epsilon_{it} \quad (1.1)$$

where y_{it} is some outcome variable of interest (e.g., the log number of first-time URM students). Standard errors are clustered by institution.

Specifically, P_{it} takes on a value of “1” in any academic year when an institution’s matriculating class is affected by the policy. For example, if an institution adopted the policy during the 2015-2016 academic year for the incoming class of 2020, it is first indicated as test-optional in the 2016-2017 academic year. When y_{it} denotes graduation outcomes, P_{it} is lagged by 6 periods to appropriately reflect how cohorts of students are affected by the policy.²⁰

The term α_i represents institution fixed effects while λ_t represents academic year

²⁰Graduation data on IPEDS reflects cohorts enrolling 6 years prior to data reporting. For example, graduation rate data from 2016-2017 reflects the cohort entering in 2010-2011. So, P_{it} is also lagged by 6 periods when the outcome variable is the 4-year graduation rate.

fixed effects. Institution fixed effects should control for time-invariant factors that may confound the estimates of the policy’s effects. The vector \mathbf{X}_{it} includes all of the control variables. Some of these variables, such as institutional grants per FTE, can control for some additional policy changes that may accompany the adoption of the test-optional policy, such as financial aid expansion, and therefore confound its effects.²¹

Section 1.5.2 explores the potential issue of heterogeneity in the policy’s effects across colleges’ admissions in selectivity. This subsection splits test-optional institutions into two groups: “selective” and “highly selective” institutions. To perform this analysis, this study estimates the following regression:

$$y_{it} = \beta P_{it} + \delta(S_i \cdot P_{it}) + \mathbf{X}'_{it}\gamma + \alpha_i + \lambda_t + \epsilon_{it}. \quad (1.2)$$

It includes an interaction term $S_i \cdot P_{it}$ where S_i takes on a value of “1” if institution i is highly selective in admissions. Thus, $\beta + \delta$ represents the effects of the policy for these types of institutions while β alone captures the the effects for selective institutions. δ captures the difference in effects between these two types of institutions. The vector \mathbf{X} includes the same covariates as in equation (1.1) plus those interacted with S_i . α_i and λ_t are the same fixed effects from equation (1.1), although the latter is also interacted with S_i .

Finally, Section 1.5.3 compares the policy’s effects on first-time URM enrollment outcomes between colleges dropping the test-optional policy early in the panel versus later. To distinguish between early and later adopters of this policy, this study estimates

²¹As previously discussed, the adoption of the test-optional policy has been noted to be paired with financial aid expansion. To my knowledge, there are no other known types of policy changes that have accompanied the adoption of this policy.

the following specification:

$$y_{it} = \beta_1(P_{it} \cdot E_i) + \beta_2(P_{it} \cdot L_i) + \mathbf{X}'_{it}\boldsymbol{\gamma} + \alpha_i + \lambda_t + \epsilon_{it}. \quad (1.3)$$

Here, y_{it} corresponds to either the log number of first-time URM students or the fraction of first-time students of a URM background. E_i is a binary variable that takes on a value of “1” if test-optional institution i adopted the policy early within the time frame. Similarly, L_i takes on a value of “1” if a test-optional institution adopted the policy later. Both E_i and L_i always equal “0” for all test-requiring institutions. Hence, the test of $\beta_1 = \beta_2$ can indicate whether the policy’s effects between early and late-adopters are distinguishable.

1.4.2 Dynamic Treatment Effects & Identifying Assumptions

This study estimates the dynamic effects of the test-optional policy on the outcome variables of interest to check for pre-treatment trends and the common trends assumption. Specifically, this study estimates the following regression specification:

$$y_{it} = \sum_{\tau=-7}^{-2} \delta_{\tau} D_{it} + \sum_{\tau=0}^{7} \mu_{\tau} D_{it} + \mathbf{X}'_{it}\boldsymbol{\gamma} + \alpha_i + \lambda_t + \epsilon_{it}. \quad (1.4)$$

Similar to equation (1.1), y_{it} represents an outcome variable, \mathbf{X}_{it} represents the same vector of controls, and α_i and λ_t represent institution and year fixed effects, respectively. δ_{τ} and μ_{τ} represent the leading and lagging coefficients across event time τ , i.e., the relative number of periods to adoption of the test-optional policy. This equation omits $\tau = -1$, so the leads and lags represent the dynamic effects of the policy relative to one period prior to adoption. The leading and lagging periods are symmetric across

event time.²² However, since the adoption of the test-optional policy is staggered, the test-optional institutions from the sample are unbalanced in these periods.

1.5 Results

1.5.1 Average effects among all colleges

Figures 1.2-1.5 contain the plotted dynamic effect estimates across event time for all outcome variables. Table 1.2 contains the p -values from the joint tests on the leading coefficients across each outcome (i.e., these test $\delta_\tau = 0$ for $-7 \leq \tau \leq -2$). These tests fail to find evidence that at least one leading coefficient is statistically different from 0. That said, the common trends assumption is presumed to hold for all outcome variables.

Recent literature suggests the coefficients for the leads and lags may be contaminated by effects from other periods when the treatment is staggered (Sun and Abraham, 2021). Consequently, this issue may invalidate the joint test on the leads. To address the above issue, this study re-estimates the dynamic treatment effects using alternative estimators proposed by de Chaisemartin and D’Haultfœuille (2021) that are robust to treatment heterogeneity. Further discussion on these dynamic effects can be found in Appendix A.4, which shows that the results of an analogous test also presume that the common trends assumption holds.

Columns (1) and (2) of Table 1.3 contain the point estimates of the TWFE coefficient from equation (1.1) for outcomes on first-time URM enrollment. It indicates that the policy raises the fraction of first-time students from a URM background by about 0.0188 points (i.e., on average, test-optional colleges experience a 23.5% increase from the pre-treatment average). Also, these estimates show that the number of first-time URM

²²The chosen number of periods is roughly based on the median number of leading and lagging periods across treated units from the sample.

students rises by 12.5%. This suggests that the increased fraction can be primarily attributed to an increased inflow of first-time URM students enrolling at these institutions rather than a drop in the overall volume of enrollees.²³

Although the average effects of the policy on URM enrollment are positive, the effects for each test-optional college in the sample may vary by size. This study investigates further by calculating the difference in first-time URM enrollment outcomes between two periods after and before policy adoption for all test-optional colleges within the sample. Figure 1.6 displays these differences for the share of first-time students from a URM background within a histogram. It illustrates positive differences among the majority of test-optional colleges in the sample. Figure 1.7 also indicates the abundance of positive differences for the log number of first-time URM students enrolling at these colleges. These figures collectively suggest that most test-optional colleges are likely to implement equitable consideration between submitters and non-submitters of test scores and that they experience warming effects from URM students.

All in all, these results suggest that the policy is, to a small extent, effective at bolstering racial diversity at liberal arts colleges. As a possibility, however, these patterns may be driven by an overall increase in the volume of enrolling students, regardless of their race. In that case, the increase in the shares of URM students may be a byproduct of that goal. However, Appendix A.6 discusses the effects of the policy on the logged number of first-time non-URM students, and it suggests that this is not likely to be the case.

Finally, columns (3) and (4) of Table 1.3 displays the test-optional policy's effects on the graduation outcomes of URM students. The point estimate for each outcome is relatively small (e.g., a 0.02 point decrease in the 4-year URM graduation rate) and

²³As an exercise, Appendix A.5 shows that the policy has little impact on the enrollment outcomes of non-first-time URM students.

statistically insignificant. These results suggests that the policy has a negligible impact on the 4-year and 6-year graduation rates for URM students. So, on average, racial diversity improvements resulting from the policy may not diminish beyond the first year of college. However, further discussion of these effects across school selectivity is provided in the next subsection.

The TWFE approach to estimating these effects have a few caveats. First, colleges and universities may voluntarily elect to drop their test-requirement. Thus, the “treatment” takes the form of an endogenous policy change. This could be an issue if colleges with relatively lower shares of URM students *self-select* themselves into this treatment. However, the treated and control colleges from the sample have comparable and statistically indistinguishable shares of URM students on their campuses (i.e., the difference is 0.004 with $p = .62$).²⁴ Also, as shown in Appendix A.2, the estimates are robust to the exclusion of control variables. Finally, this study re-estimates equation (1.1) using a propensity-trimmed sample and finds that the resulting point estimates are comparable to that of the full sample. The logistic regression used to construct this sample suggest that the pre-treatment share of URM students on campus is not a significant predictor for adopting the test-optional policy. Some further discussion behind this propensity trimmed sample can be found in Appendix A.3.

Alternatively, the endogeneity of policy adoption may stem from colleges within each state following neighboring institutions in dropping the test requirement. This study collapses the panel to the state level and conducts analogous state-level analysis to investigate whether this phenomenon is evident among colleges in the sample. Specifically, it estimates the impact of the share of students attending a test-optional college within each state on state-level freshman URM enrollment outcomes. The results of this ex-

²⁴This share corresponds to the overall fraction of students on campus that are from a URM background, regardless of freshman status.

ercise can be found in Appendix A.9. They indicate that the estimates of the share of treated students on state-level URM enrollment outcomes are noisy. In short, there is little evidence that colleges' decision to drop the test requirement reflects the actions of neighboring in-state institutions.

Second, recent papers show that the TWFE estimator with staggered treatment is a weighted sum of all possible 2×2 difference-in-differences estimators (Goodman-Bacon, 2021). However, some of these weights may be negative in the presence of treatment heterogeneity which could render the estimator uninterpretable (de Chaisemartin and D'Haultfœuille, 2020). The TWFE estimators for the URM enrollment outcomes (e.g., the fraction of first-time students that are URM) are composed of 536 ATTs, of which 5.6% of them receive negative weights. These weights sum up to about -0.012. On the other hand, the TWFE estimators associated with the graduation rates contain zero negative weights. Appendix A.4 discusses estimates for the URM enrollment outcomes using an alternative estimator that is robust to treatment heterogeneity, and it suggests that negative weighting is not a significant issue.

1.5.2 Differences in Admissions Selectivity

The analyses, thus far, encompass institutions regardless of their degree of selectivity in admissions. Highly selective institutions tend to place a relatively higher weight on test scores within their admissions processes (Marin and Horn, 2008). Therefore, as a possibility, the effects of the test-optional policy may differ at colleges considered to be more selective than others. So, the effects of the policy may be more salient at the most selective colleges. To investigate this possibility, this study distinguishes institutions that are “selective” and “highly selective” (i.e., are considered more selective and most selective by the USNWR). The sample contains 54 selective institutions and 95 highly

selective institutions of which 44.4% and 45.3% are test-optional, respectively. The summary statistics for these two sub-samples can be found in Tables A1.2 and A1.3.²⁵

Equation (1.2) is estimated to distinguish the effects of the policy across institutional selectivity. The point estimates are displayed in Table 1.4 across all outcome variables. Interestingly, among the URM enrollment outcomes, the point estimates for selective colleges among the URM enrollment outcomes are higher than that for highly selective colleges. However, the differences between these two sub-groups are insignificant. Appendix A.6 shows that neither of these sub-groups is likely to be adopting the test-optional policy to increase overall enrollment, regardless of students' racial background. Although the effect of the policy on the first-time enrollment volume at selective colleges is significant, it is much smaller than the effect on first-time URM enrollment.

The point estimates of the 4-year and 6-year URM graduation rates are negative and significant among selective colleges (e.g., the 4-year graduation rate drops by 0.0678 points, or 12.8% from the pre-treatment level). On the other hand, the point estimates for highly selective colleges are small and insignificant, and in fact, the differences between the two sub-samples are significant. In other words, the policy led to a decline in the URM graduation rates at selective colleges. At these institutions, students benefitting from the policy may be less likely to graduate.²⁶ Consequently, the graduation rates for URM students decrease as a result of the policy.

Further discussion on these graduation rate effects is provided in Section 1.6. Appendix A.8 shows that test-optional admissions have little impact on graduation rates for non-URM students (e.g., White and Asian students) at both selective and highly

²⁵The mean URM graduation rates for selective colleges are larger than that of highly selective colleges, with the difference for 6-year graduation rate being statistically significant. However, this difference within the propensity-trimmed sample from Appendix A.3 is slightly smaller but statistically insignificant. The results of the analyses in that section is very comparable to the ones discussed below (i.e., under the full sample).

²⁶This could be backed by the fact that the mean pre-treatment admission rate of selective colleges is higher than that of highly selective colleges by about 22%.

selective colleges.

1.5.3 Timing of Policy Adoption

As discussed, test-optional institutions within the sample adopted the policy throughout different years between 2002-2003 and 2019-2020. However, the effects of the policy on URM enrollment may vary across these institutions by the timing of treatment, i.e., whether they dropped the test requirement relatively early within the panel's time frame or later. As one possibility, the warming effects on enrollment may be more prevalent among colleges adopting the policy relatively early in the panel. Specifically, if colleges drop the test requirement early within an environment where there are few or no other test-optional colleges, they would have an easier time attracting prospective URM students seeking test-optional admissions. Consequently, colleges that drop the test requirement relatively later in this time frame may experience little policy effects on URM enrollment because prospective students seeking test-optional admissions already have an alternative: the early adopting colleges. Thus, the policy's impact on racial diversity at late-adopting colleges may be negligible if they are crowded out by early adopters.

This study distinguishes institutions that adopted the policy *early* in the panel versus *later* by estimating equation (1.3) with first-time URM enrollment outcome variables. It considers early adopters as colleges that dropped the test requirement anytime up to 2010, i.e., the midpoint of the time frame. All other test-optional colleges that dropped their test requirement after 2010 are considered to be late adopters.

Table 1.5 contains the point equations of equation (2.3), where each column corresponds to an outcome variable. The point estimates show that the average effects on first-time URM enrollment outcomes between institutions adopting the policy during the first versus the second half of the time frame are comparable and statistically indistin-

guishable. All in all, these results indicate that there is no strong evidence that the policy's effects differ between early and late adopters.²⁷

Appendix A.2 reproduces the analysis of Table 1.5 without the use of control variables, and it illustrates similar patterns. Appendix A.6 shows that neither early (i.e., adopting through 2007) nor late adopters experienced a significant increase in enrolling first-time students, regardless of race.

1.6 Discussion

This study finds that the test-optional policy effectively bolsters racial diversity at liberal arts colleges. It also finds that the policy has little impact on the URM graduation rates at highly selective colleges. But it finds evidence that the policy led to a decline in graduation rates at relatively less-selective colleges. Finally, this study finds that the URM enrollment effects among colleges dropping the test requirement early in the panel versus later are comparable and indistinguishable.

Given the sign of the point estimates, these results are consistent with a warming effect. While most residential liberal arts colleges today are not necessarily highly selective or elitist, individuals are inclined to equate a residential liberal arts education with an "elite" form of higher education (Astin, 1999). Therefore, these institutions tend to be associated with the perception of prestige/selectivity. Thus, their adoption of this policy may send a signal to prospective URM students that they are trying to facilitate a welcoming environment for them.

These results suggest that the mismatch effect may be present at selective colleges versus highly selective ones. Indeed, the most selective institutions tend to draw higher

²⁷This study considers other ways to define early adopters, such as those that drop the test requirement through 2007 (i.e., the first 25% of adopters). However, under this alternative threshold, the effects from early and late adopters are also statistically indistinguishable.

shares of applicants from the top quintile of their graduating high school cohort (Bound et. al, 2009). Thus, high-ability students who presumably face greater probabilities of retention and successful graduation may sort themselves to these institutions, even within a test-optional admissions regime. However, the relationship between graduation rates and the test-optional policy is unlikely to be causal. These estimates may encompass other unobservable policies and phenomena that affect retention, such as the availability of advising and outreach at highly selective colleges or the lack thereof at relatively less-selective institutions for students benefitting from the policy.²⁸ Therefore, there is no certainty whether the test-optional policy alone led to the decline in the URM graduation rates at selective colleges and whether mismatch is a significant issue within test-optional admissions.

As suggested by the results, these test-optional colleges may have implemented an enhanced, equitable admissions process for submitters and non-submitters of test scores. However, the outcome of interests used in this analysis reflects enrollment. Although enrollment effects can pose implications for admissions, it would do so, albeit to a limited extent. Unfortunately, IPEDS does not disaggregate its admissions data by demographics. Thus, richer data is necessary to fully understand how the policy affects the admissions processes for URM students at these institutions.

Furthermore, as discussed in Section 1.2, some test-optional colleges place greater weight on inequitable admissions criteria, such as high school rigor. Therefore, not all URM applicants may be benefitting from the policy. However, the estimated effects of the policy on URM enrollment outcomes suggest that this channel is dominated by warming effects or the prevalence of equitable admissions within the sample.

²⁸The estimates of the policy's effects on URM enrollment may not be exempt from this issue as well. As discussed in Footnote 17, however, the test-optional policy is likely paired with financial aid expansion or shifts in weighting toward college preparatory courses, which would have been captured by the control variables.

Certainly, the policy’s effects on liberal arts colleges could yield implications on its beneficiaries’ post-graduation returns. A recent study, Carnevale et. al (2020), indicates that the median long-run (40-year) net present value (NPV) returns of liberal arts colleges is approximately \$918,000, i.e., almost \$200,000 higher than the median for all types of 4-year colleges and universities.²⁹ Also, it finds that the returns to the most selective liberal arts colleges (i.e., most of the institutions within the “highly-selective” category of the sample) are higher at \$1,135,000. Since the policy does not seem to diminish the graduation rates of URM students at these institutions, students benefitting from the policy could experience significant returns on investment and possibly upward social mobility.³⁰ On the other hand, the negative impact of the policy on the URM graduation rates at relatively less-selective institutions may be problematic. Carnevale et. al (2020) also shows that high graduation rates are correlated with high investment returns.³¹ Therefore, students who benefit from the test-optional policy but face lower probabilities of graduation attainment will not be able to realize these gains.

Some questions about the test-optional policy remain for future investigation. By its design, the policy’s effects should be salient to students from financially disadvantaged backgrounds. Although IPEDS data on such students is limited, an analysis of this policy on *socioeconomic* diversity in admission, enrollment, and graduation would be informative in understanding its effects.³² Next, other types of institutions, such as

²⁹Other types of colleges & universities, as designated by the Carnegie Classification, include doctoral-granting universities and special-focus schools.

³⁰Carnevale et. al (2020), contains the estimated 40-year returns for a significant amount of individual liberal arts colleges. By using these figures, this current study finds comparable trends within its sample. The median 40-year return to investment among all 149 institutions is about \$949,000 which also exceeds the average for all types of institutions. However, the median is higher at \$999,000 compared to \$882,500 for selective institutions.

³¹This finding is also consistent with patterns observed within this current study’s sample. The correlation between the estimated 40-year return to investment of colleges included in the sample and the 2020 graduation rates is 0.691.

³²Although IPEDS contains data on the number of first-year Pell grant recipients at each college, the availability of that variable is constrained to a small number of periods.

doctoral-granting universities, are increasingly dropping their test requirement, so future studies may be able to assess the impact of the policy on racial or even socioeconomic diversity at these types of schools. The analyses of public institutions may especially be of interest since many of these systematically target a large and relatively diverse array of students, so the impact of the policy on these schools may be pronounced.³³ Finally, future analyses can shed light on how the test-optional policies shift enrollment patterns between different types of institutions (e.g., private versus public or selective versus less-selective institutions).³⁴

Racial and economic gaps in college enrollment and attainment continue to persist, despite the relatively high returns to post-secondary education and significant student aid efforts. This study primarily contributes to the discussion of how policies can bolster racial diversity within colleges and universities and considers the test-optional policy as a measure that can achieve this goal. Interestingly, many colleges and universities that dropped the test requirement during the COVID-19 pandemic are either prolonging their admission regimes or making them permanent.³⁵ As the costs and benefits of this policy come to light, colleges and universities will become better informed on whether to maintain their admissions regime. With the ongoing prevalence of test-optional admissions among colleges and universities, knowledge of these costs and benefits is becoming more relevant than ever.

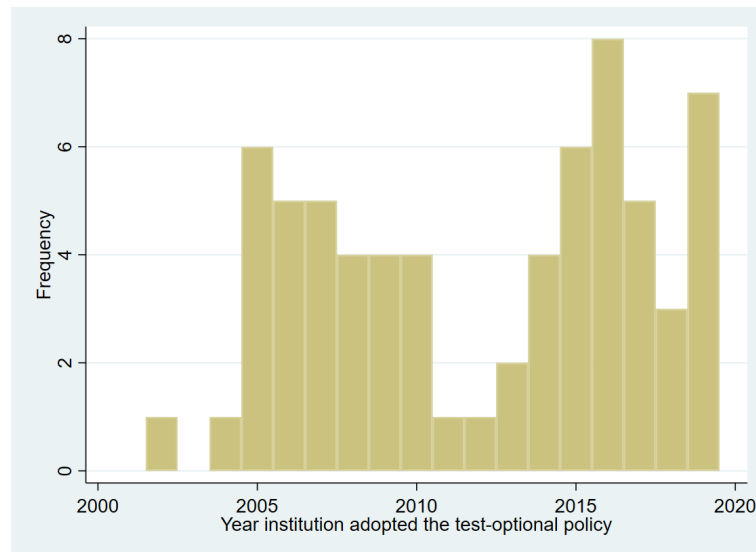
³³For example, the 1960 California Master Plan for Education stipulates that the University of California (UC) system, which elected to permanently drop its test requirement in 2020, targets the top 12.5% of California high school students. In fact, within the past few years, the UC system has been accommodating roughly 18% of California high school students, and they consider expanding further to match the growing demand for admission. (Source: <https://www.latimes.com/california/story/2021-07-27/california-is-failing-to-meet-demand-for-uc-admission-why-its-a-crisis>)

³⁴For example, Hinrichs (2012) and Backes (2012) are studies that assess the impact of an admission policy (i.e., affirmative action bans) on the enrollment patterns of URM students. They find that the policy shifted the fraction of URM students attending more selective colleges to less-selective ones. They posit that the ban also led to an increase in enrollment of these students at public 2-year colleges, but they find little empirical evidence of that occurring, mostly owed to data limitations.

³⁵For example, Occidental College, which dropped its test requirement at the onset of the pandemic, notes on its website that it has no plans to return to test-requiring admissions in the future.

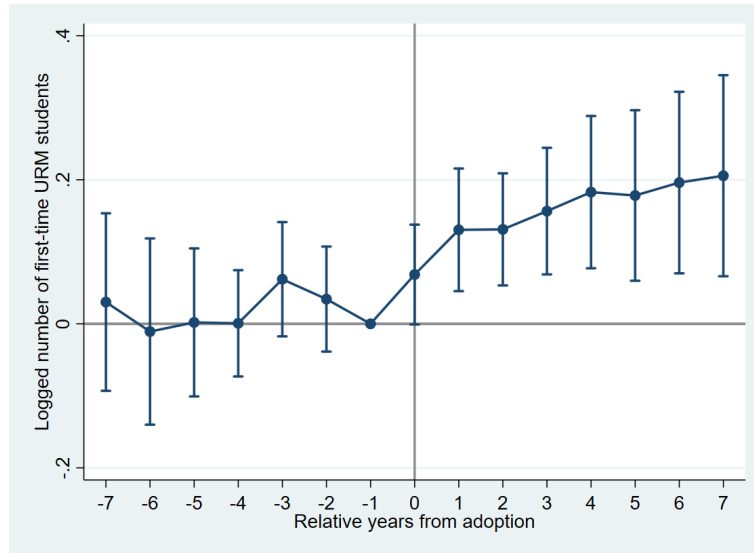
1.7 Figures

Figure 1.1: Distribution of Policy Adoption Year



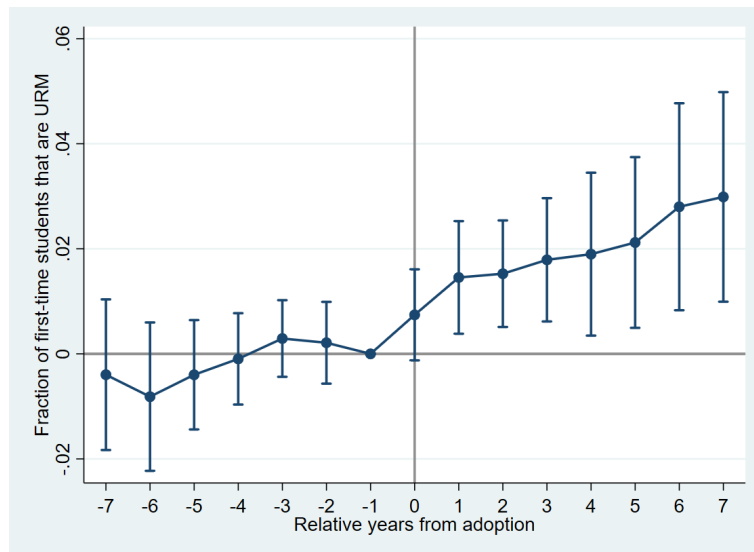
Notes: This figure displays the distribution of the years in which test-optional institutions from the sample adopted their admissions policy.

Figure 1.2



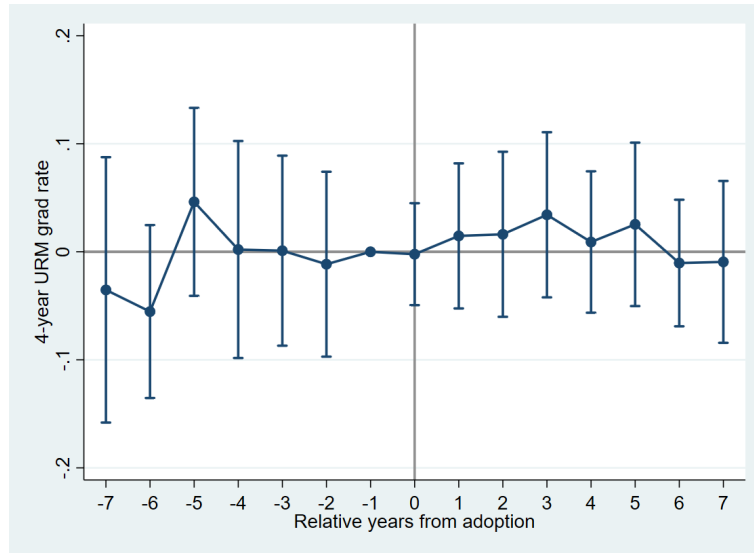
Notes: This figure illustrate the dynamic effects on the logged number of first-time URM students. Coefficient estimates from equation (1.4) are plotted across event time (i.e., years relative to the test-optional policy taking effect). They are represented by the blue dots. The accompanying bands represent the 95% confidence intervals of these coefficients. Figures 1.3-1.5 are arranged similarly, but they reflect different outcome variables.

Figure 1.3



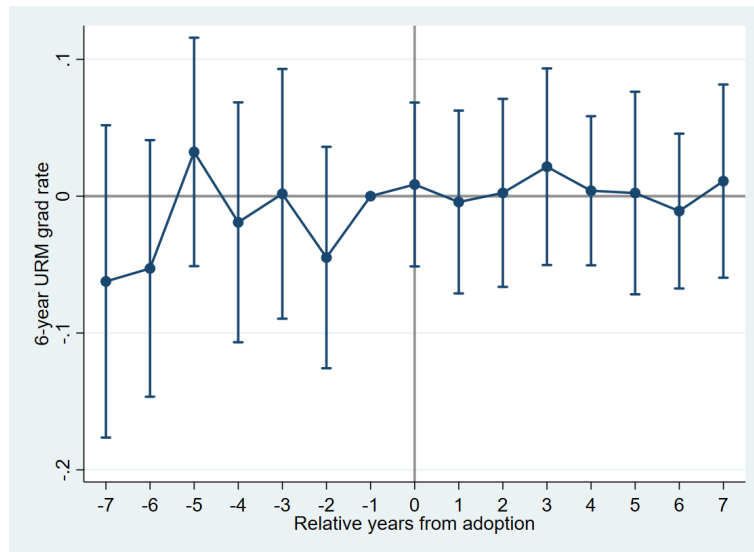
Notes: This figure illustrates the estimated dynamic effects of the test-optional policy on the fraction of first-time students enrolling at liberal arts colleges that are of URM status.

Figure 1.4



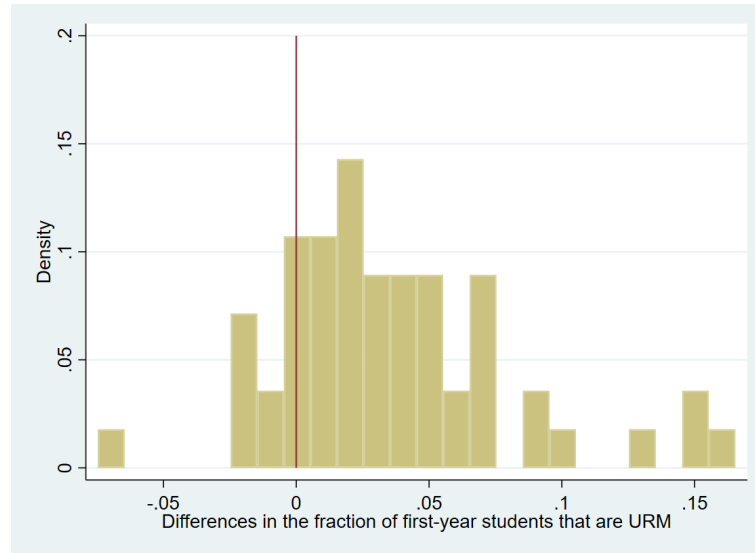
Notes: This figure illustrates the estimated dynamic effects of the test-optional policy on the 4-year graduation rate for URM students.

Figure 1.5



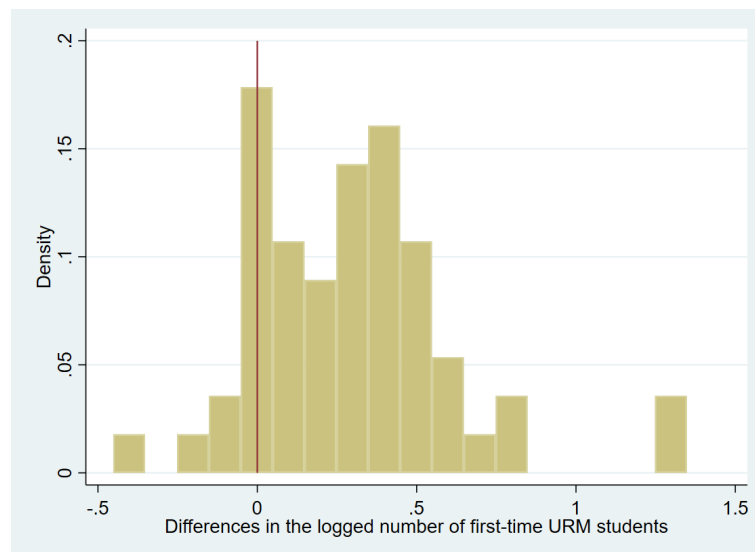
Notes: This figure illustrates estimated dynamic effects of the test-optional policy on the 6-year graduation rate for URM students.

Figure 1.6



Notes: For each test-optional college in the sample, this study calculates the difference in the share of first-time students from a URM background two periods after and before policy adoption (i.e., $Y_i^{2+} - Y_i^{2-}$). This figure is a histogram for these differences, where the vertical red line corresponds to “0” difference.

Figure 1.7



Notes: For each test-optional college in the sample, this study also calculates the difference in the logged number of first-time URM students background two periods after and before policy adoption (i.e., $Y_i^{2+} - Y_i^{2-}$). This figure is a histogram for these differences, where the vertical red line corresponds to “0” difference.

1.8 Tables

Table 1.1: Summary Statistics

	Test- optional (1)	Test- requiring (2)	p-value of diff. (3)
First-time URM students	31.00 (20.39)	29.87 (25.90)	0.77
Fraction of first-time students that are URM	0.08 (0.05)	0.08 (0.05)	0.74
4-year URM grad rate	0.53 (0.19)	0.52 (0.24)	0.78
6-year URM grad rate	0.63 (0.16)	0.60 (0.23)	0.48
Tuition & Fees	30,290 (5,547)	27,957 (7,729)	0.03
Full-time enrollment	1,489 (597.9)	1,372 (714.9)	0.28
Institutional grants per FTE	9.70 (4.73)	20.00 (43.03)	0.03
E & R expenditures per FTE	28.41 (17.99)	69.92 (166.56)	0.03
College prep courses not considered	0.03 (0.17)	0.02 (0.16)	0.84
College prep courses recommended	0.36 (0.48)	0.62 (0.49)	0.00
College prep courses required	0.60 (0.49)	0.34 (0.48)	0.00
Observations	67	82	–

Notes: Columns (1) and (2) contain summary statistics for test-optional and test-requiring institutions using the 2001-2002 (i.e., pre-treatment) observations. Standard deviations are in parenthesis. Column (3) contains the p -values from the difference means between these two groups. These p -values are clustered by institution. None of the variables are logged. Therefore, the means for tuition & fees, E & R expenditures per FTE, and institutional grants per FTE are in terms of 2019 dollars. The sample size varies slightly across each variable due to non-reporting.

Table 1.2: Results of Joint Test of Leads

	<i>p</i> -value for joint F-test
a) Logged number of first-time URM Students	0.588
b) Fraction of first-time students that are URM	0.823
c) 4-year URM graduation rate	0.385
d) 6-year URM graduation rate	0.276

Notes: This table displays the *p*-values for the joint test in leading coefficients from equation (1.4) across all outcome variables of interest. The point estimates of the leads and lags are displayed in Figures 1.2-1.5.

Table 1.3: Primary Results

	Logged Number of first-time URM students (1)	Fraction of first-time students that are URM (2)	4-year URM grad rate (3)	4-year URM grad rate (4)
Test-optional	0.125** (0.0419)	0.0188** (0.00628)	-2.32e-05 (0.0142)	0.00450 (0.0131)
Observations	2,820	2,824	2,783	2,812

Notes: This table displays the results from estimating equation (1.1) on all outcome variables, and it solely reports the TWFE estimate of the test-optional policy. Columns (3) and (4) correspond to graduation outcomes, so the regressions for those use a lagged treatment indicator (i.e., $P_{i,t-6}$). One star indicates a 5% significance level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

Table 1.4: Effects of the Policy across Selectivity

	Logged Number of first-time URM students (1)	Fraction of first-time students that are URM (2)	4-year URM grad rate (3)	6-year URM grad rate (4)
Test-optional	0.154* (0.0690)	0.0247* (0.0121)	-0.0678** (0.0153)	-0.0476** (0.0149)
Highly Selective \times Test-optional	-0.0422 (0.0831)	-0.00792 (0.0139)	0.0820** (0.0211)	0.0662** (0.0200)
Observations	2,827	2,831	2,790	2,819

Notes: This table displays the point estimate from equation (1.2), which distinguishes institutions by their degree of admission selectivity. The row labeled with “Test-optional” contains the point estimate for the coefficient of the treatment indicator of having the policy in place. The row labeled “Highly selective \times Test-optional” contains the point estimate for the coefficient of the interaction term between the treatment indicator and another indicator for being highly selective in admissions. Each column corresponds to an outcome variable. Similar to Table 1, the point estimates for the coefficients of the control variables are not reported here. One star indicates a 5% level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

Table 1.5: Effects of the Policy by Adoption Timing

	Logged Number of freshman URM students (1)	Fraction of freshman students that are URM (2)
Early Adopter	0.117 (0.0597)	0.0208* (0.0102)
Late Adopter	0.133** (0.0516)	0.0164* (0.00814)
<i>p</i> -value	0.830	0.751
Observations	2,820	2,824

Notes: This table displays the point estimate from equation (1.3), which distinguishes the effects of the policy by adoption timing. The row labeled “Early adopter” corresponds to the estimated effect of the policy on institutions that dropped the test requirement early. Similarly, the row labeled “Later adopter” corresponds to the effect on institutions adopting the policy later. Each column corresponds to a first-time URM outcome variable. The *p*-values on the bottom row reflect the test of $\beta_1 = \beta_2$. One star indicates a 5% level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

Chapter 2

When Is Discrimination Unfair?

2.1 Introduction

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

To study this question, we use a vignette-based survey experiment on Amazon's Mechanical Turk (MTurk) to measure people's assessments of the fairness of race-based hiring decisions. The vignettes illustrate canonical examples of statistical and taste-based discrimination, with both Black and White recipients of discrimination (*discriminatees*). In addition, the scenarios have varying levels of *justifiability*, i.e., varying motivations for the discriminatory act which we expect will make the actions more or less socially acceptable. The goals of our analysis are, first, to measure the effects of three types of factors on the perceived fairness of a discriminatory act in a broad sample of Americans:

the characteristics of the respondent; the motivation for discrimination (e.g., tastes versus statistical); and the identity of the discriminatee (Black versus White). Second, we assess the consistency of four pre-registered models of perceived fairness with the patterns we observe. Finally, we provide a simple, non-preregistered, two-group interpretive framework that provides a convenient summary of all our empirical results.

Our main findings are as follows. First, subjects' self-identified political leanings have large effects on their overall acceptance of discriminatory actions, with conservatives being much more accepting of the discriminatory actions we depict than moderates and liberals. Second, regardless of their political leanings, our respondents care about the detailed motivations behind a discriminatory action (holding the act's consequences constant). Specifically, while the presence of taste-related versus statistical factors does not reliably predict subjects' fairness assessments, other aspects of the discriminator's motivations have robust and sizable effects. For example, discrimination by employers is seen as substantially less fair when it is based on the employer's own tastes than on the tastes of the employer's customers. Similarly, statistical discrimination is seen as less fair when it is based on low-quality information about relative group productivity, compared to higher-quality information. Notably, the effects of these motivational factors on perceived fairness are very similar across all political groups, and the effects do not depend on the race of the discriminatee.

Third, our moderate and liberal respondents exhibit a strong *discriminatee race effect*: they disapprove more of anti-Black than anti-White discrimination. This effect is absent among conservatives, who rate the discriminatory acts we depict as slightly more fair than unfair, regardless of the discriminatee's race. Fourth, among the four models of perceived fairness we evaluate – (simple) utilitarianism, race-blind rules (RBRs), racial in-group bias, and belief-based utilitarianism (BBU) – the latter two are inconsistent with some major empirical patterns in our data. Fifth, all three political groups in our

sample –liberals, moderates, and conservatives—exhibit a strong desire to apply race-blind rules when comparing the fairness of different discriminatory actions. Only liberals and moderates, however, exhibit utilitarian preferences (which assign higher fairness ratings to actions that shift income from high-to low-income groups).

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

Finally, we notice that – unlike Group 1 – Group 2 subjects (who are all moderate or liberal) could face a conflict between the two fairness criteria (utilitarianism and race-blind rules) they care about: Objecting more strongly to anti-Black than anti-White discrimination for utilitarian reasons may not feel race-blind. To assess how Group 2 makes this tradeoff, we leverage the fact that a large share of our subjects experiences a switch in the race of the person being discriminated against during the experiment. Under the assumption that subjects only become aware of their desire to make race-blind fairness assessments after they are exposed to a discriminatee of a second race, we are able to use random assignment of our *race* treatments to estimate that Group 2’s fairness assessments place roughly equal weight on these two criteria when they conflict.

Our paper connects to a literature in labor and personnel economics that uses models of fairness to interpret the effects of pay inequality on effort, job performance and satisfaction, wage satisfaction, and quits (Charness and Kuhn 2007; Abeler et al. 2010;

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

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

¹In 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.

²One 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.

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

Finally, our analysis relates to ongoing debates among both economists and legal scholars about which forms of discrimination are more 'egregious' than others (and there-

³Considering 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.

fore perhaps more deserving of policy remedies or legal sanctions.) For example, in a recent review article, Bertrand and Duflo (2017) provide the following description of a common view among economists:

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

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

Section 2.2 of the paper describes our survey design, data collection, and sample characteristics. Section 2.3 presents some basic facts about fairness perceptions: How

⁴The word “equally” is not present in Bertrand and Duflo’s text; we have inserted it to convey what we believe is their meaning.

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

do perceptions vary with respondent characteristics, survey treatments, and interactions between the two? Section 2.4 describes four simple, preregistered models of fairness and compares their implications to subjects’ aggregate response patterns. It shows, – among other things— that our Group 2 respondents (*Utilitarians*) care about two criteria — utilitarianism and race-blindness – that sometimes conflict. Section 2.5 then uses within-subject variation in the discriminatee’s race to estimate the relative weight these *Utilitarian* subjects place on the two criteria they care about. Section 2.6 concludes.

2.2 Design and Implementation

2.2.1 Survey Structure

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

Next, our subjects are randomly assigned to read four vignettes in which an employer, “Michael” (or “Andrew”) makes a hiring decision between a White and a Black applicant.⁶ These scenarios are designed to represent canonical examples of taste-based and statistical discrimination. We define taste-based discrimination as a decision that

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

is based on *someone's* racial animus, and distinguish two forms: *Less-justifiable* taste-based discrimination is based on the employer's *own distaste* for people of a particular race.⁷ *More-justifiable* taste-based discrimination occurs when the employer accommodates his *customers'* distastes for a particular race. Statistical discrimination, on the other hand, is based on differences in expected job performance. *Less-justifiable* statistical discrimination is based on low-quality information about the relative performance of two racial groups; we frame this as a non-quantitative statement from a single, non-expert source (a 'neighbor') about problems others experienced when employing White or Black employees, such as lateness and lack of attention to detail. *More-justifiable* statistical discrimination is based on "reliable statistics" from "a large and experienced network of local business owners" who frequently hire for the same type of opening as in the scenario.⁸ In both cases, the 'justifiability' rankings of the more detailed reasons for the action are based on our own priors regarding how respondents would react. After reading each vignette, the respondent is asked to rate the fairness of the employer's hiring decision on a scale from 1 to 7. As in Alesina et al. (2020, 2021), these questions have no material consequences.⁹

The four scenarios encountered by each respondent are presented in two Stages. In Stage 1, subjects were assigned with equal probability to one of the four possible treatment combinations: SW, TW, SB, and TB, where S and T represent statistical and taste-based discrimination, and W and B indicate the race of the discriminatee. Within

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

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

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

Stage 1, the subjects encounter the less- and more-justifiable versions of discrimination in random order. In Stage 2, subjects were randomly assigned to one of the three treatment combinations they did not encounter in Stage 1, and again encountered the more-versus less-justifiable forms in random order.¹⁰ Thus, as illustrated in Appendix B1.2, two thirds of the respondents encountered a switch in the discriminatee’s race, and two thirds encountered a switch between taste-based and statistical discrimination.

Our survey concludes with Stage 3, which asks all subjects the same questions. First, in an open-text question, we remind respondents of the final scenario they encountered and invite them to explain the fairness assessment they made. Next, we use the following question to elicit subjects’ assessment of Black people’s relative economic opportunities (BRO):

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

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

Much less / Less / A Little Less / Roughly equal / A little more / More / Much More opportunity than White people.

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

2.2.2 Scenario 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

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

introduce this Stage, we first tell subjects they will encounter two scenarios which share many common elements but contain some differences; we also say that the differences have been underscored to make them easier to pick out. The subjects then read and assess the *less* or *more* justifiable forms of the Taste discrimination scenario with a Black discriminatee in random order. The *less* justifiable form of taste discrimination is motivated by the employer's own tastes:

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

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

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

He has conducted focus groups with a substantial share of the people who frequent his business. Many of these customers tell Michael that they do not like interacting with Black people and would be hesitant about continuing to support his business if he employed them. Michael himself is just as happy to interact with Black workers as with workers of other races.

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

After each scenario, the respondent is asked to "indicate the extent to which you thought

that Michael’s hiring decision was fair” on a seven-point scale, where 1 was “very unfair”, 4 was “neither fair nor unfair”, and 7 was “very fair”.

As noted, in Stage 2, respondents encountered two more scenarios in which either the *race* of the discriminatee (Black or White), the *motivation* for the discrimination (Tastes versus Statistical) or both of these were different from Stage 1.¹¹ White scenarios were identical to Black scenarios except that the races of the discriminator and the discriminatee are reversed. As noted, *less* justifiable statistical discrimination was based on low-quality information (hearsay from a single, uninformed source) about relative group productivity, and *more* justifiable statistical discrimination was based on higher-quality information (quantitative information from substantial sample of other employers). The exact wording of these and all our scenarios is provided in Appendix B.1.1.

2.2.3 Implementation and Representativeness

On September 21, 2020, we pre-registered our survey design and procedures, and posted a pre-analysis plan. Our survey was administered to a sample of MTurk workers between September 22 and October 6, 2020. Subjects were given one hour to complete the survey and were informed that we expected the task to take about 15 minutes. Conditional on completing the entire survey, subjects were paid \$5.¹² A few measures were taken to improve the accuracy and representativeness of the responses. First, respondents were required to have a U.S. address. Second, to further discourage foreign workers from participating, the survey was launched during U.S. Pacific daylight hours on weekdays.

¹¹Exposing subjects to four scenarios (rather than one) has three main benefits. First, it gives us more fairness assessments, while preserving the option to use only each subject’s first treatment for pure cross-subject comparisons. Second, as illustrated in Section 2.5.1, it allows us to test for a specific but plausible form of experimenter demand effects. Third, as illustrated in Section 2.5.3, it allows us to assess the relative importance subjects assign to utilitarian, versus race-blind fairness criteria when those criteria conflict.

¹²In 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.

Third, MTurk workers were required to have a 95 percent approval rating to discourage robots (i.e., automated responses). Fourth, the survey included a CAPTCHA question to further discourage robots. Finally, respondents were exposed to each vignette for at least 30 seconds before being allowed to submit their fairness assessment. In all, we received 779 responses; during data cleaning we dropped 137 of these, leaving us with a final count of 642 responses in our analysis sample.¹³

In Appendix B.2, we present summary statistics of our survey respondents and compare them to adults in the 2019 American Community Survey (ACS) and the 2020 General Social Survey. Compared to the ACS, our sample of MTurkers is quite regionally representative, a little more male, and a little more likely to be either White or Black. Our respondents are also considerably better educated and much more likely to be between 25 and 44 years of age than U.S. adults in general. For the most part, these are well known features of the MTurk population.¹⁴ Comparing our subjects' political orientations to the GSS is more difficult because—despite the similarity of the survey questions—the middle category differs between the two surveys: “moderate” in our case versus “moderate, middle of the road” in the GSS.¹⁵ Ignoring this difference in phrasing, it would appear that our MTurk respondents are politically more ‘extreme’ than GSS respondents, with more candidates selecting the two extreme categories and far fewer selecting the middle one. However, GSS respondents could be attracted to the attracted

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

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

¹⁵Since the ACS does not collect information on political opinions or affiliations, we are forced to use the GSS (with its much smaller sample size) to assess the political representativeness of our population. Our political party preference question is not comparable to the GSS's, but (with the exception of this middle category) our political leaning question is identical to the GSS's (see Table B2.2 for details).

to the ‘middle of the road’ label.¹⁶ In sum, our MTurk-based sample differs from the U.S. population in substantial and mostly well-known ways. In Appendices B.9 and B.10 we estimate the implications of these differences by re-weighting our main results to match the American Community Survey and General Social Survey respectively. The results are very similar.

2.2.4 Question Order Effects

In all multi-part surveys, but especially in contexts like ours where framing and experimenter demand effects might play a large role, the order in which respondents encounter different questions could have large effects on the respondents’ answers. We address this issue in detail in Appendix B.3, which shows that question order effects are absent from our survey in two distinct senses. First, as shown in Appendix B.3.1, there is no time trend in fairness assessments across the four scenarios encountered by each respondent: Respondents become neither more nor less accepting of discrimination as they are asked additional questions about it.¹⁷ Second, the order in which respondents encounter the Tastes versus Statistical and the *more* versus *less* justified scenarios does not affect their fairness ratings on subsequent scenarios. In Appendices B.3.2 and B.3.3, 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

¹⁶In addition to this possible difference in variance, there is also some suggestion that, on average, MTurkers are somewhat more liberal than GSS respondents. Almost identical shares of MTurkers and GSS respondents choose some degree of conservative leaning (ranging from slight to extreme), but many more MTurkers choose some liberal leaning (47.3 versus 30.2 percent). This difference could reflect the relative youth of MTurkers.

¹⁷Recall 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.

were equal but opposite in sign to subjects who switched in the opposite direction. Finally, for both the type of discrimination and the *justifiability* treatments we show that within-subject, between-subject and pooled fairness regression estimates are statistically indistinguishable from each other.¹⁸

The one treatment that does, however, affect subjects' subsequent fairness assessments is the *race* of the discriminatee. As we document in Appendix B.3.4, our respondents' Stage 2 fairness assessments depend on the *race* treatment they encountered in Stage 1. In the next two Sections of the paper (2.3 and 2.4), we will eliminate the influence of these order effects by relying only on data from Stage 1 of the survey: There is no within-subject variation in the *race* treatment during Stage 1 (or during Stage 2), because discriminatee race only varies between the experiment's two Stages. In Section 2.5, 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.

2.3 Some Facts

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

¹⁸Within-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.

¹⁹There 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

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

2.3.1 How Does Perceived Fairness Vary with Respondents' Characteristics?

In Figure 2.1, we show how the mean perceived fairness of discriminatory acts varies with respondents' characteristics. To maximize our sample size for these initial comparisons, we pool responses across all four treatment combinations (SW, TW, SB, and TB) as well as the more- and less-justifiable versions of both types of discrimination. In short, Figure 2.1 shows that our subjects' mean fairness assessments do not vary significantly with their age or race. However, women viewed the discriminatory acts as slightly less fair than men. Somewhat surprisingly (to us), respondents' fairness assessments were positively related to their education levels; we explore this correlation in Appendix B.4 and show that higher levels of education mostly reflect a higher 'set point' for all fairness assessments in the following senses. First, regardless of political leaning, more-educated individuals rate *all* the scenarios they encounter as more fair than less-educated individuals. Furthermore, in contrast to political leaning—which has strong effects on how our subjects respond to some of our treatments—highly-educated subjects respond to all our treatments in very similar ways to less-educated subjects.²⁰

Finally, Figure 2.1 shows that respondents' political leaning is strongly related to the perceived fairness of discriminatory acts. Self-described conservative respondents

the very first of the four scenarios each person encountered, although the standards errors are somewhat higher. See for example Figure B5.1.1, which replicates Figure 2.1 using first-scenario data only.

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

perceive these actions to be fairer than both moderates and liberals (e.g., $p = .000$ for conservatives versus liberals).²¹ Mean fairness assessments across U.S. political party preferences (e.g., Democrats versus Republicans) exhibit similar patterns.²² Since the Independent group could include people with both extreme right- and left-wing orientations, we use conservative-liberal leaning rather than party affiliation to categorize respondents' political preferences in the remainder of the paper.²³

2.3.2 Treatment Effects

In Figure 2.2, we compare the fairness assessments of subjects who were exposed to the Tastes versus Statistical treatments, and to the more versus less justifiable forms of each. As in Section 2.3.1. we pool both of the Stage 1 scenarios encountered by each worker and cluster our standard errors by respondent. To simplify the presentation, we also pool the Black and White treatments.²⁴ To facilitate interpretation here and throughout the paper, we report all fairness assessments on a scale from -3 (“very unfair”) to 3 (“very fair”), where 0 was labeled in the survey as “neither fair nor unfair.”²⁵ The standard deviation of fairness assessments in Stage 1 is 1.657 across respondents, 0.961 within respondents, and 1.915 overall.

According to Figure 2.2, the average respondent sees no meaningful distinction between the fairness of the statistical versus taste-based scenarios in our survey ($p = .971$). Conditioning on whether discrimination is taste-based or statistical, however, subjects

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

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

²³Of the respondents who identify themselves as “Independent” within our sample, about 8.43% suggest they are either “extremely liberal” or “extremely conservative.” Furthermore, all the results by political party are very similar, with occasional non-monotonicities similar to Figure 2.1(e), where Independents appear to be to the left of Democrats.

²⁴Figure 2.3 shows that the effects of justifiability are virtually identical for White versus Black discriminatees.

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

view the less justifiable form of either taste-based or statistical discrimination as less fair than the more justifiable form ($p = .000$ in both cases), confirming our expectations. To illustrate the size of these differentials, we first remark that an average respondent did not view the more-justifiable forms of either statistical or taste-based discrimination (high quality information; accommodating the tastes of others) as unfair at all: the mean fairness ratings of these actions were in the “somewhat fair” range with small standard errors.²⁶ In contrast, the less 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 2.3 we turn our attention to the race treatment – i.e. the race of the person who was discriminated against. Motivated by Figure 2.2 (which shows no difference between the Statistical and Tastes treatments) we now pool these treatments but continue to distinguish between their more- versus less-justifiable forms. In the sample as a whole, Figure 2.3 shows that respondents view the same discriminatory acts more negatively when they are directed at Black than at White job applicants. We call this phenomenon the *Discriminatee Race Effect* (DRE). The DRE shown in Figure 2.3 is substantial in magnitude, amounting to about 0.5 fairness units or 0.263 standard deviations, and highly statistically significant ($p = .002$ and $.000$ within the less versus more justifiable forms of discrimination, respectively).

²⁶The 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.

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

While the effects of the race treatment shown in Figure 2.3 are interesting, these effects may not be the same for all types of respondents. For example, Black respondents might react more negatively than White respondents to discrimination against Black job applicants. To explore this issue, Figure 2.4 presents separate estimates of the discriminatee race effect for respondents of different races. Unfortunately, our samples of both Black and Other racial groups are too small to precisely estimate a discriminatee race effect within either group. The point estimates for these groups however suggest that both groups respond to the race of the discriminatee in much the same way as White respondents do.²⁷ In sum, Figure 2.4 underscores the fact that the discriminatee race effect in our data – i.e., the tendency to see discrimination against Black people as less acceptable than discrimination against White people—is driven primarily by our White respondents, who comprise about 78 percent of the sample. Thus, while we continue to estimate all our results on the full sample of MTurk respondents in the remainder of the 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 2.5 presents separate estimates of the discriminatee race effect by the respondent’s political leaning. These reveal a clear difference: the discriminatee race effect is stronger among moderate and liberal respondents than in the sample as a whole, but is absent among conservatives. Conservatives view discrimination against (fictitious and identically qualified) Black and White job appli-

²⁷Interestingly, Figure 2.4 indicates that the Other group views discrimination relatively harshly. However, there is little indication of a discriminatee race effect for this group of respondents ($p = .506$) and the point estimates themselves are imprecise.

cants the same way: as more fair than unfair.²⁸ A final striking finding from Figure 2.5 is the strong similarity in both the levels of fairness rankings and in the discriminatee race effects between self-described moderate and liberal respondents. Later in the paper (starting in Section 2.4.4) we exploit this fact to simplify our analysis by comparing just two political groups—conservatives versus moderates/liberals.

2.4 Assessing Four Models of Fairness

This Section describes four simple models of how subjects might evaluate the fairness of discriminatory actions: (simple) utilitarianism, racial in-group bias, race-blind rules (RBR), and belief-based utilitarianism (BBU). For each model, we compare its predictions with the main empirical patterns in our data and show that two of the models – racial in-group bias and BBU – are inconsistent with some key patterns in our data. After examining subjects’ open-text responses for clues that might explain these inconsistencies, we then propose a two-group framework with three fairness criteria that does a better job of accounting for the facts we have documented. As in Section 2.3, our analysis only uses data from Stage 1 of the experiment to ensure that race treatment order effects cannot affect our conclusions.

2.4.1 (Simple) Utilitarianism

In general, utilitarian models of fairness share two main features, the first of which is that fairness depends on outcomes, not on intentions or justifications. Since our Tastes vs. Statistical and less-versus-more justifiable treatments refer to the *reasons* for the employer’s actions, and since the consequences of the employer’s actions — i.e. which

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

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

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

Turning to the evidence on simple utilitarianism, Figure 2.3 has already shown that respondents in general *do* view discrimination against Black applicants more harshly than discrimination against White applicants. That said, Figure 2.5 showed that this tendency was confined to moderates and liberals: Conservatives do not consider race when assessing the fairness of discriminatory actions. We conclude that our (simple) utilitarian model is consistent with moderates' and liberals' fairness assessments, but

is not consistent with conservatives' fairness statements.²⁹ Utilitarianism also cannot account for the large justifiability effects on perceived fairness that are documented in Figures 2.2 and 2.3.

2.4.2 Racial in-group bias

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

As Figure 2.4 has already shown, we do not have the statistical power to test these predictions for the respondents in our Black or Other racial categories.³¹ Our evidence for White respondents, however, is strongly inconsistent with racial in-group bias: As a group, White respondents view discrimination against Black people as substantially *less* fair than discrimination against White people. Interestingly, when we focus our attention

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

³⁰Related (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.

³¹In 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.

on the subset of White respondents who identify as conservative, this strong rejection of in-group bias no longer holds: As shown in Figure B5.2.1, White conservatives rate discrimination against Black people as 0.405 units *more* fair than discrimination against White people. This difference is however not statistically significant at conventional levels ($p = .134$). Overall, we conclude that racial in-group bias model does not provide a useful lens for understanding the main fairness ratings patterns we have documented.

2.4.3 Race-Blind Rules (RBR)

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

Applying these ideas to our experiment, an RBR model of fairness would – unlike the previous two models – allow the fairness of a discriminatory action to depend on the intentions behind it: Did the act serve to indulge the employer’s personal racial animus, or to protect his business from retaliation by racist customers? Did the employer do his due diligence before relying on statistical information in hiring, or did he take hearsay-based shortcut? Further, assuming the respondent has an implicit set of rules defining

³²Intentions 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.

which of the above motivations are fairer than others, she should apply those rules in a race-blind way. For example, if using low-quality statistical information is x units less fair than using high-quality information, x should be the same regardless of the race of the discriminatee.

The fairness ratings of our respondents are consistent with the use of race-blind rules (RBRs) in at least three ways.³³ First, the effects of our *justifiability* treatments in Figure 2.2 strongly support the idea that respondents care about the employer’s motivation for discriminating against a job applicant. Importantly, our experimental design ensures that the *justifiability* effects in Figure 2.2 hold the consequences of the discriminatory action constant: While the material consequences of not being hired could, for example, vary with the discriminatee’s race (because of differences in outside labor market options), notice that Figure 2.2 varies only the *reasons* for not being hired: discriminatee race is balanced between the motivation and justifiability treatments due to random assignment.

Second and more strikingly, Figure 2.3 shows that our respondents penalized the less-justifiable forms of discrimination by the same amount (relative to the more justifiable forms), *regardless of the race of the discriminatee*: (-0.953 versus -0.898 fairness points for White versus Black discriminatees respectively, with $p = .679$ for a test of equality). Third, a similar test shows that this stability to discriminatee race also applies to the Taste/Statistical fairness differential – it is essentially zero for both Black and White discriminatees.³⁴ A final, remarkable feature of our respondents’ apparent adherence to race-blind rules is that it applies just as strongly on both sides of the U.S. political divide. We show this explicitly in Figure 2.6, which shows that respondents ranked the

³³In Sections 2.5.2 and 2.5.3, we will present a third piece of evidence supporting the RBR model that applies only to moderate and liberal respondents. Specifically, we will argue that the order effects for the Black treatment (which are present only for moderate and liberal respondents) suggest that these respondents prefer to maintain a form of consistency across race in their fairness assessments.

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

relative fairness of *more versus less* justifiable forms of discrimination almost identically, irrespective of their political leaning.

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

2.4.4 Belief-Based Utilitarianism (BBU)

In Section 2.4.1 we ruled out (simple) utilitarianism among conservative respondents because those respondents did not object more strongly to anti-Black than to anti-White discrimination, even though Black job applicants, on average, have lower incomes. This fact, however, does not rule out the possibility that conservatives are motivated by a different form of utilitarianism, which we label *belief-based utilitarianism* (BBU).³⁵ Under BBU, respondents still value redistribution from more- to less-advantaged groups, but they use a different and possibly subjective metric of relative advantage to guide their fairness evaluations.³⁶ From a modeling perspective, BBU is an appealing hypothesis because it would allow us to explain a key empirical difference between conservatives and other respondents—conservatives do not exhibit a discriminatee-race effect—in a

³⁵BBU 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.

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

straightforward way: Both conservatives and other respondents are in fact utilitarians (i.e. they prefer to favor a disadvantaged group) but they simply have different beliefs about who is disadvantaged.

Evidence that is consistent with BBU is presented in Figure 2.7, which draws on the BRO question in Stage 3 of our survey. This question asked the respondents to rate Black people’s relative economic opportunities in the United States on a seven-point scale, running from “much less opportunity” (minus 3 in Figure 2.7) to “much more opportunity” (plus 3 in Figure 2.7). Figure 7 shows that the respondents’ BRO ratings differ dramatically by their political orientation: While liberals have a mean BRO of -1.374 ($p = .000$), conservatives’ mean of -0.206 is insignificantly different from zero ($p = .089$) with moderates in between. This suggests that conservatives’ belief that Black and White people have roughly equal opportunities has the potential to account for their observed fairness ratings, which –like their BRO ratings—are statistically the same for discrimination against Black versus White job applicants.³⁷

To assess whether BRO differences can actually account for the partisan gap in fairness assessments, panel (a) of Figure 2.8 shows respondents’ fairness ratings for anti-Black discrimination by BRO categories, separately for conservatives and moderates/liberals.³⁸ If BRO accounts for the large partisan gap, we should see little or no partisan gap *within* the BRO categories: Instead, the partisan gap should be explained by the higher mean

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

³⁸Starting in this subsection, we combine moderates and liberals into a single group to simplify the presentation and preserve statistical power. Interested readers can view a version of Figure 2.8 with all political groups in Appendix B.5.3; all the qualitative results discussed below are similar for moderates and liberals individually, as well as the combined group.

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 2.8 shows that there are large political gaps in the perceived fairness of discriminating against Black people, even when we condition on BRO. These political gaps are particularly stark at the bottom of the BRO distribution: While moderates and liberals who think that "Black people have much less opportunity than White people" (BRO = -3) are strongly opposed to anti-Black discrimination, conservatives with the same belief are, on average, *accepting* of anti-Black discrimination (with a mean fairness rating of about +0.5). This partisan gap at the bottom of the BRO distribution is highly statistically significant. Within subjects who have BRO levels of -3, and within subjects who have BRO levels of -2, the partisan gap is significant at $p = .000$.

A third and even more surprising piece of evidence against the "BRO hypothesis" emerges from panel (b) of Figure 2.8, which replicates panel (a) for discrimination against White job applicants. Consistent with a large explanatory role for BRO in peoples' fairness assessments, this Figure shows an effect of BRO on the perceived fairness of discrimination that is essentially invariant to political orientation: the coefficients are .257 ($p = .094$) and .265 ($p = .000$) for conservatives and moderates/liberals respectively. However, for both political groups the direction of this effect is the opposite of what the BRO hypothesis would predict: According to the BRO hypothesis, higher levels of Black people's perceived relative opportunities should make discrimination against White people less acceptable. Instead, the perception that Black people have equal or more economic opportunities than White people – which is held by 36.9 percent of

our subjects—is associated with a greater tolerance of (hypothetical) acts of anti-White discrimination.

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

2.4.5 What Motivates Respondents Who Say Discrimination is Fair?

To try to make sense of the unexpected findings in Figure 2.8, we first observe that a large subset of our respondents (when classified by beliefs and political orientation) exhibit a common pattern of fairness assessments: they assign roughly equal, non-negative fairness to both anti-Black *and* anti-White discriminatory acts. In Figure 2.8, this group of respondents includes all self-identified conservatives, *plus* the moderates and liberals with $BRO \geq 0$. Together, these respondents (henceforth Group 1) represent 48.9 percent of all respondents. In Figure 2.8 they are indicated by the hollow circle markers.

To understand what types of fairness criteria might account for these respondents’ fairness assessments, we then turned to our respondents’ open-text explanations of the last scenario they encountered. As described in Appendix B.6, we manually classified these explanations into three broad categories, separately for respondents who rated

discrimination as “unfair” or “very unfair”, versus respondents who said it was “fair” or “very fair”. For the latter group, the most common type of response was some variation of ‘*the business must thrive*’, such as:

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

“A business wants [to] retain customers and high profits. So give them what they expect. Hiring what people prefer is reasonable.”

“Andrew needs to do what is best for his business. If he hired the black worker, he’d lose money and perhaps even go out of business.”

“To ensure the success of his business, Michael should do everything possible to do so. Is it racist? I don’t think so. Michael is free to hire whoever he wants.”

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

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

“It’s his company he can hire whoever he choses [sic]. He does not have to give an answer to anyone or share his hiring views. He can choose what is best at any time.”

“It’s his business. He doesn’t need to justify to me any of his hiring practices. . . . Seriously, he does not need to justify his hiring choices.”

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

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

“I’ve never had a problem with this as long as every business owner is allowed to do it; I don’t feel comfortable in all-black establishments so just I don’t go to them. They’d be uncomfortable, I’d be uncomfortable, just let them have their thing. What’s the problem?”

“Employers are allowed to hire whoever they see as best for the job.”

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

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

2.4.6 An Interpretive Framework that ‘Works’

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

In our proposed framework, there are two main groups of respondents. Group 1 (the “Business Rights Advocates”, accounting for 48.9% of all respondents) includes all conservatives, plus moderates and liberals with $BRO \geq 0$. These respondents are, on average, accepting of all the discriminatory actions we depict, regardless of the race of the discriminatee. Members of Group 1 justify these fairness assessments as protecting or raising profits and preserving employer rights. Members of Group 1 do not appear to be motivated by any utilitarian concerns (either ‘simple’ or ‘belief-based’). Based on our findings in Section 2.4.3, however, they share the widespread support across the political spectrum for race-blind procedural fairness (RBRs).

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

2.5 Reconciling Conflicting Fairness Criteria: Utilitarianism versus RBRs

In the previous Section, we proposed a two-group interpretive framework in which one group of respondents (Group 1 – *Business Rights Advocates*) cares only about race-blind rules (RBRs) and individual decision rights, while Group 2 (*Utilitarians*) cares

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

only about RBRs and utilitarian welfare criteria. In this framework, there is a clear potential for conflict between Group 2's two main fairness objectives: Rating anti-Black discrimination more harshly than similar acts of anti-White discrimination may not feel race-blind.⁴⁰ In this Section, we use the race treatment order effects documented in Section 2.2.4 to estimate the relative weight that members of Group 2 place on these two fairness criteria. Our identifying assumption is that respondents are not aware of their desire to make race-blind fairness assessments until they encounter a discriminatee from a second racial group, i.e. until they encounter a switch in the race treatment.

To accomplish this goal, we proceed in three steps. First, we document that the race treatment order effects described in Section 2.2.4 are present in Group 2 but absent in Group 1. Second, we use a simple model of reporting behavior, combined with random assignment of the *race* treatment to interpret Group 2's order effects as a compromise between utilitarianism versus RBRs, and to estimate the relative weight Group 2 places on those two criteria when they conflict. Finally, for readers who may be skeptical of our *ex-post* Group 1 - Group 2 dichotomy, we replicate the preceding steps for the categories used in Section 2.4 of the paper: conservatives versus [moderates + liberals]. The results are very similar.

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

In Section 2.2.4 we demonstrated the existence of race treatment order effects in our entire sample of respondents: Their Stage 2 fairness assessments depend on the race treatment they encountered in Stage 1. In Appendix B.8.1, we show that there are no

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

such order effects for Group 1: Regardless of the discriminatee race they encountered in Stage 1, members of Group 1 view discrimination as a little more fair than unfair (about +0.5 on a scale from -3 to +3) in Stage 2. Respondents from Group 2, on the other hand, exhibit a more pronounced version of the order effects we saw in the sample as a whole. Specifically, Group 2's Stage 2 fairness assessments of anti-Black discrimination are much milder if they encountered a White discriminatee (as compared to a Black discriminatee) in Stage 1 ($p = .009$).⁴¹ Motivated by this fact, we restrict our attention to Group 2 respondents (Utilitarians) in the following sub-section, where we interpret Group 2's race treatment order effects as a compromise between the values they place on both utilitarianism and race-blindness.

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

As noted above, subjects who value both utilitarianism and race blindness (e.g. members of Group 2) face a conflict when they experience a switch in the race treatment. For example, in Stage 2, a White-to-Black treatment switcher needs to choose between assigning the same fairness rating they assigned to a White discriminatee in Stage 1 (race blindness), versus respecting their utilitarian desire to object more strenuously to anti-Black than anti-White discrimination. Subjects who do not experience race treatment changes do not face this conflict.

To model this idea, we make the following assumptions:

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

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

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

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

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

$$B_i^2 = \alpha B_i^* + (1 - \alpha)W_i^1 \quad (2.1)$$

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

To address this unobservability problem we take advantage of the fact our race treatments are randomly assigned. Thus, while B_i^* is not observed for W-to-B switchers (and W_i^* is not observed for B-to-W switchers), their sample means \bar{B}^* and \bar{W}^* in any fixed population (such as Group 2) are observed for both groups of switchers from the mean Stage 1 responses of the subjects in their population who were randomly assigned to the other race treatment. We can therefore write:

$$\bar{B}^2 = \alpha \bar{B}^* + (1 - \alpha)\bar{W}^* \quad (2.2)$$

⁴²Subjects' 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.

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

Similarly, for B-to-W switchers.

$$\bar{W}^2 = \alpha \bar{W}^* + (1 - \alpha) \bar{B}^* \quad (2.3)$$

After restricting our sample to Group 2 respondents, Equations (2.2) and (2.3) can then be (separately) solved for α , yielding $\alpha = 0.49$ for the White-to-Black switchers and $\alpha = 0.68$ for the Black-to-White switchers.⁴³ Thus, W-to-B switchers' Stage 2 ratings of anti-Black discrimination place almost equal weight on race-blindness and utilitarianism. B-to-W switchers, on the other hand, act as if they place slightly more weight on utilitarianism than on race-blindness. The 95% percent confidence intervals for α are [0.243,0.800] and [0.405, 1.075] for W-to-B and B-to-W switchers, respectively. Thus, we cannot reject equal weight on both objectives ($\alpha = 0.5$) for either type of switcher. For W-to-B switchers, we can reject both $\alpha = 0$ and $\alpha = 1$, indicating strictly positive weight on both objectives. For B-to-W switchers, we reject $\alpha = 0$ but not $\alpha = 1$.

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

2.5.3 Replicating the Analysis by Political Orientation

In Appendix B.8.2, we replicate the preceding analysis, splitting the sample by political orientation (conservatives versus moderates/liberals) with very similar results. Specif-

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

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

2.6 Robustness

One potential concern about the external validity of our results is the fact that our data were collected in September and October 2020, following a summer of civil unrest related to the murder of George Floyd on May 25, 2020. Together, these events led to a mainstream conversation on systemic racism in the U.S.; it seems reasonable to ask whether these events may have primed our respondents to answer our questions in unusual ways. To check for this possibility, Figure B2.1 presents online search trends for related keywords, including *Black Lives Matter* and *racism* during the spring and summer of 2020. These trends show that searches for these terms had diminished dramatically by the time of our survey, suggesting that this type of priming may not have been a significant issue for our respondents.

One striking result of our analysis is the large magnitude, statistical significance, and stability of the *justifiability* treatments, documented in Section 2.4.3: Respondents of all political orientations penalized the less-justifiable forms of discrimination (relative to more-justifiable forms) by the same amount, irrespective of the discriminatee's race.

⁴⁴See Appendix B.8.2 for the details underlying these calculations.

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 how they rank the fairness of these two types of scenarios, with respect to each other. To address this issue, Appendix B.5.1 replicates Tables 1 and 2 using only the first individual scenario each respondent encountered. The results are almost identical to our main estimates, suggesting that subjects' desires to maintain a consistent ranking of the two types of scenarios are not responsible for this finding.

Figure 2.8 illustrated some strong and unexpected relationships among subjects' beliefs about relative opportunities (BROs), subjects' political orientation, the race of the discriminatee, and subjects' fairness assessments. Together, these relationships were starkly inconsistent with the belief-based utilitarian model of fairness. To probe the robustness of these results to the fact that Figure 2.8 combines moderates and liberals into a single group, Figure B5.3 replicates Figure 2.8, showing separate results for moderates versus liberals. Consistent with other results in the paper, these two groups exhibit very similar response patterns, both of them differing substantially from conservatives' patterns.

In Section 2.5 of the paper, we used *race* treatment order effects to estimate the relative weight a subset of our subjects assign to utilitarian and race-blind fairness criteria. These order effects could, however, be caused by a particular form of experimenter demand effects that seems quite plausible in our context. To see this, consider the following possibility: If respondents encounter the Black treatment in Stage 1, they assume that we (the experimenters) are either moderate or liberal. Then – to please us – the respondents provide Stage 1 fairness assessments that are typical for moderates and liberals (i.e. discrimination against Black applicants is unfair, and more unfair than discrimina-

tion against White applicants). On the other hand, if respondents encounter the White treatment in Stage 1, they assume the experimenters are conservative and provide Stage 1 answers that are typical for conservatives (i.e., discrimination against both Black and White applicants is neutral or fair). Finally, respondents who encounter a change in the *race* treatment between Stages 1 and 2 update their priors to become uncertain about the experimenters' politics and moderate their fairness assessments accordingly. Together, these patterns could account for exactly the type of race treatment order effects we observe, where (for example) respondents' Stage 2 opposition to anti-Black discrimination is reduced if they encountered anti-White discrimination in Stage 1.

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

In Appendix B.8, we examine whether subjects' responses to these Stage 3 questions depend on the race treatments they received in Stages 1 and 2, and find no such effects: Specifically, subjects' BRO assessments, stated political party preferences, and reported left-right leaning are highly stable with respect to the race treatments they encountered earlier in the experiment. We conclude that experimenter demand effects of this type

are probably not responsible for the race treatment order effects we observe.⁴⁵ While not conclusive, the stability of subjects' Stage 3 responses to previously-encountered race treatments also suggests the experimenter demand effects are likely not responsible for the strong and robust *justifiability* effects we estimate.

A broader concern that could affect the validity of all our results is the fact that our sample of MTurk respondents was not representative of adult Americans on a number of key dimensions, including age and education (see Appendix B.2 for details). While our small sample size limits what we can do to address this issue, Appendices B.9 and B.10 replicate all our main results (Figures 2.2-2.8) two different ways. First, Appendix B.9 uses the 2019 American Community Survey to re-weight our MTurk responses by the relative prevalence of our respondents in 24 cells, defined by gender, race, education, and age. All the main patterns discussed in the paper are replicated, with one small exception: the weak positive association between BRO and the fairness of anti-Black discrimination among conservative respondents in Figure 2.8(a) becomes somewhat stronger and statistically significant. Similar to Figure 2.8, however, the slope for conservatives remains much lower than the slope for moderates/liberals. Second, Appendix B.10 replicates Figures 2.2-2.8 using weights derived from the 2020 General Social Survey (GSS) which are based only on a 7-point political leaning scale (i.e., extremely conservative, conservative, slightly conservative, moderate, slightly liberal, liberal, and extremely liberal) that is asked in a very similar way to our survey.⁴⁶ Despite significant differences

⁴⁵Additional evidence against this demand-effects hypothesis is the fact, documented in Appendix B.8.2, that race treatment order effects are absent among conservatives. For experimenter demand effects to explain our results in Section 2.5, these demand effects must only be present among moderates and liberals. In other words, moderates and liberals should want to please an experimenter they perceive as moderate or liberal, but conservatives must have no such desire to please a conservative experimenter. In contrast, the fairness reporting model in Section 2.5 has a 'built in' explanation for conservatives' lack of order effects: Conservatives do not value utilitarian objectives, so they experience no conflict between utilitarianism and the fairness criteria they care about.

⁴⁶In our GSS re-weighting exercise we ignored the difference in the wording of the middle political leaning category between the GSS and our survey. As noted, this might exaggerate the difference between the two surveys, so it should give us an upper bound on the effects of re-weighting. Because of

in the political mix of the two surveys, all the main results are replicated.⁴⁷

As documented in Appendix P, one of the main differences between the results reported in our paper and the methods proposed in our pre-analysis plan (PAP) involves how our main outcome variable (fairness) is measured. While the PAP proposed a standardized measure of fairness, we realized that a standardized fairness measure centered at the mean fairness level across all respondents and treatments would obscure meaningful cardinal information, relative to the measure we decided to use. Our measure is centered at a conceptually meaningful level – “neither fair nor unfair” – and our measure’s integer values correspond directly to the seven fairness categories respondents could choose from. We made a similar choice with respect to our measure of Black people’s relative opportunities (BRO), centering it at “roughly equal opportunities” rather than the sample mean. To see if these decisions had any effects on the main relationships established in the paper, Appendix B.11 replicates Figures 2.2-2.8 using standardized fairness and BRO measures. There is no meaningful difference.

A common critique of non-incentivized survey experiments like ours is that subjects have little incentive to answer the questions thoughtfully, leading to results that are noisy, or simply different from what the same person might offer if they took more time to think about the question. We took a number of precautions to prevent this (see Section 2.2.3), but it remains possible that many of our respondents gave careless answers that differ from what a thoughtful person would choose. To assess this possibility, Appendix B.12 replicates all our main results (Figures 2.2-2.8) for ‘thoughtful’ respondents only, where ‘thoughtful’ is defined as taking more than the median amount of time to complete the survey. No meaningful differences were evident.

the small size of the MTurk and GSS samples, we did not re-weight our MTurk sample to mimic GSS demographic characteristics; attempts to do this yielded extreme and imprecise weights. The ACS does not ask questions about political orientation or party preference.

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

A final important concern – which affects virtually all tests of statistical hypotheses – is the extent to which the hypotheses were selected after a preliminary analysis of the data. To address this concern, we posted a registered pre-analysis plan (PAP) before launching our survey. The relationships between the analyses proposed in the PAP and the hypotheses tested in our survey are described in detail in Appendix P. Briefly, Appendices P1-P3 together comprise a “populated PAP” which reports the results of the exact tests specified in the PAP. Appendix P4 summarizes the relationship between the PAP and the paper. In a little more detail, Appendix P4 shows that the following key analyses in the paper were declared in advance: all the descriptive “facts” presented in Section 2.3; all four theoretical models of discrimination described in Sections 2.4.1-2.4.4 and the main tests thereof (the models’ names have changed slightly); the possibility of question order effects (especially for the race treatment); and the idea of using question order effects to learn about respondents’ preferences for race-blindness (see Appendix P2.5). Appendix P4 also describes the five most important ways in which our main analyses in the paper differ from the PAP. These are all relatively minor, and the populated PAP results in Appendices P1-P3 strongly suggest they do not matter. Finally, Appendix P4 notes that there are only two PAP hypothesis tests that we decided not to include in the main paper and discusses our motivations for those decisions.

2.7 Discussion

Inspired by a rapidly growing literature on the perceived fairness of pay and income inequality, and by a large literature on discrimination, we have used an MTurk survey to elicit Americans’ assessments of the fairness of canonical examples of statistical and tasted-based racial discrimination. We find, first of all, that conservative respondents are more accepting of discriminatory actions than moderate and liberals. Second, while dis-

tinguishing between statistical and taste-based discrimination has been of considerable interest to economists, whether discrimination is motivated by (someone's) tastes or by statistical reasons is not a reliable predictor of assessed fairness. Third and in contrast, respondents of all political leanings do care about other aspects of the motivation behind a discriminatory act. Specifically, our respondents agree that acting on one's *own* tastes is less fair than accommodating others' tastes, and that using imprecise or inaccurate statistical information is less fair than precise information. Indeed, respondents of all political leanings penalize these less-justifiable actions by the same amount, and do so regardless of discriminatee race, suggesting a broad area of common ground in how Americans react to different discriminatory actions. Fourth, another important partisan difference is that only moderates and liberals consider the race of the discriminatee when assessing the fairness of a discriminatory act.

Comparing the preceding findings with four pre-registered models of how respondents might make fairness assessments, we find that two of those models – in-group bias models and belief-based utilitarianism – conflict with several key patterns in our data. Using open-text data to identify an unanticipated rationale underlying some subjects' fairness assessments, we propose an *ex post* interpretive framework with two equally-sized political groups and three models of fairness – simple utilitarianism, race-blind rules (RBRs), and employer decision rights – that can account for most of the fairness patterns we observe. In this model, both political groups value using a set of race-blind rules to compare the fairness of different types of discriminatory actions. One group, who we call “Business Rights Advocates” and are mostly conservative, also value employer decision rights. The other group, “Utilitarians” is a large subset of moderates and liberals. They value utilitarian fairness criteria in addition to RBRs, but not employer decision rights. When their utilitarian and RBR objectives conflict – as when they experience a change in the experiment's *race* treatment – we estimate that members of Group 2 place about

equal weight on these two objectives.

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

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

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

better received than other policies.

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

While we have more than enough statistical power to test our pre-registered hypotheses, we also acknowledge that we lack statistical power to answer two important questions. First, does the discriminatee race effect really reverse sign (relative to the sample as a whole) among White conservatives? If White conservatives truly object more strongly to anti-White and anti-Black discrimination, this would complicate our description of conservatives, in general, as valuing race-blindness. Second, given our small sample, our data cannot shed much light on which factors explain non-White respondents' fairness

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

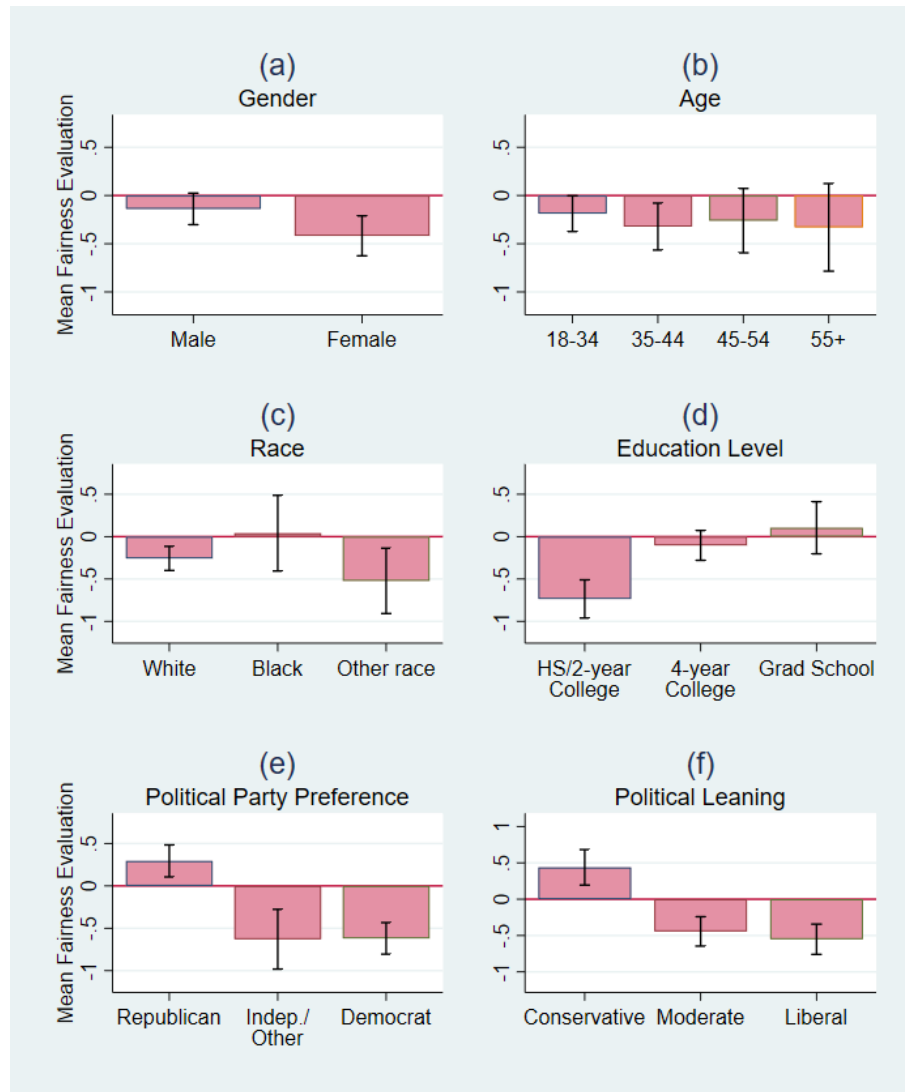
assessments. Finally, we remind readers that the fairness assessments we elicit are not necessarily the same as the *actions* our respondents might take in real-world situations similar to our scenarios. For example, a business owner might choose to accommodate the discriminatory tastes of her customers while still experiencing that action as unfair.⁵⁰ That said, given the extensive evidence that people value fairness (e.g. Card et al. 2012; Cullen and Perez-Truglia 2018; Dube et al, 2019) our paper quantifies, for the first time, the fairness *costs* our subjects associate with taking different types of discriminatory actions.

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

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

2.8 Figures

Figure 2.1: 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 is “neither fair nor unfair.” This figure is based on only Stage 1 observations. 95% confidence intervals are shown. The *p*-values below are clustered by respondent.

a) Gender

Males vs. Females = 0.037

c) Respondent race:

White vs. Black = 0.204
 Black vs. Other = 0.058
 White vs. Other = 0.197

e) Political party preference

Republican vs. Independent = 0.000
 Independent vs. Democrats = 0.961
 Democrats vs. Republicans = 0.000

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

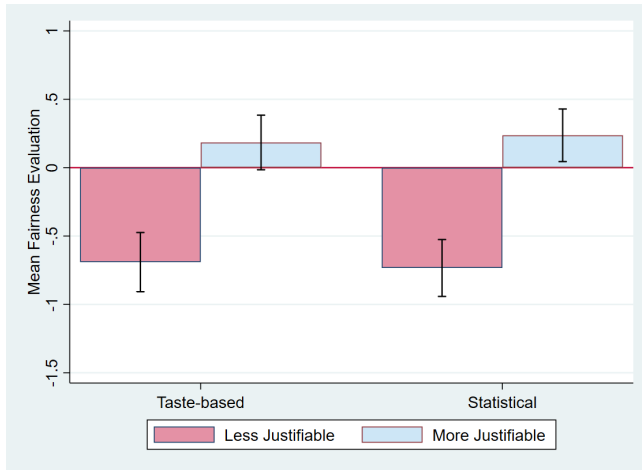
d) Education level:

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

f) Political leaning

Conservatives vs. Moderates = 0.000
 Moderates vs. Liberals = 0.463
 Liberals vs. Conservatives = 0.000

Figure 2.2: Fairness Ratings by Type of Discrimination and *Justifiability*



Less vs More Justifiable Treatment

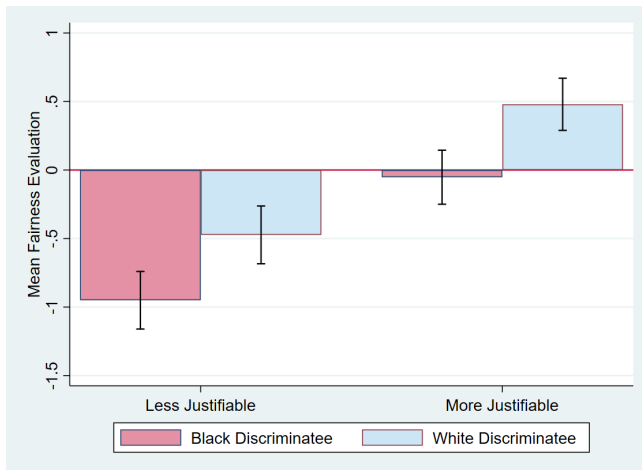
Overall: $p = .000$
 Within taste-based: $p = .000$
 Within statistical: $p = .000$

Taste vs Statistical Discrimination

Overall: $p = .971$
 Within Less-Justifiable: $p = .779$
 Within More-Justifiable: $p = .710$

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

Figure 2.3: Fairness by *Justifiability* and Discriminatee Race



Black vs White Treatment

Overall: $p = .000$
 Within taste-based: $p = .002$
 Within statistical: $p = .000$

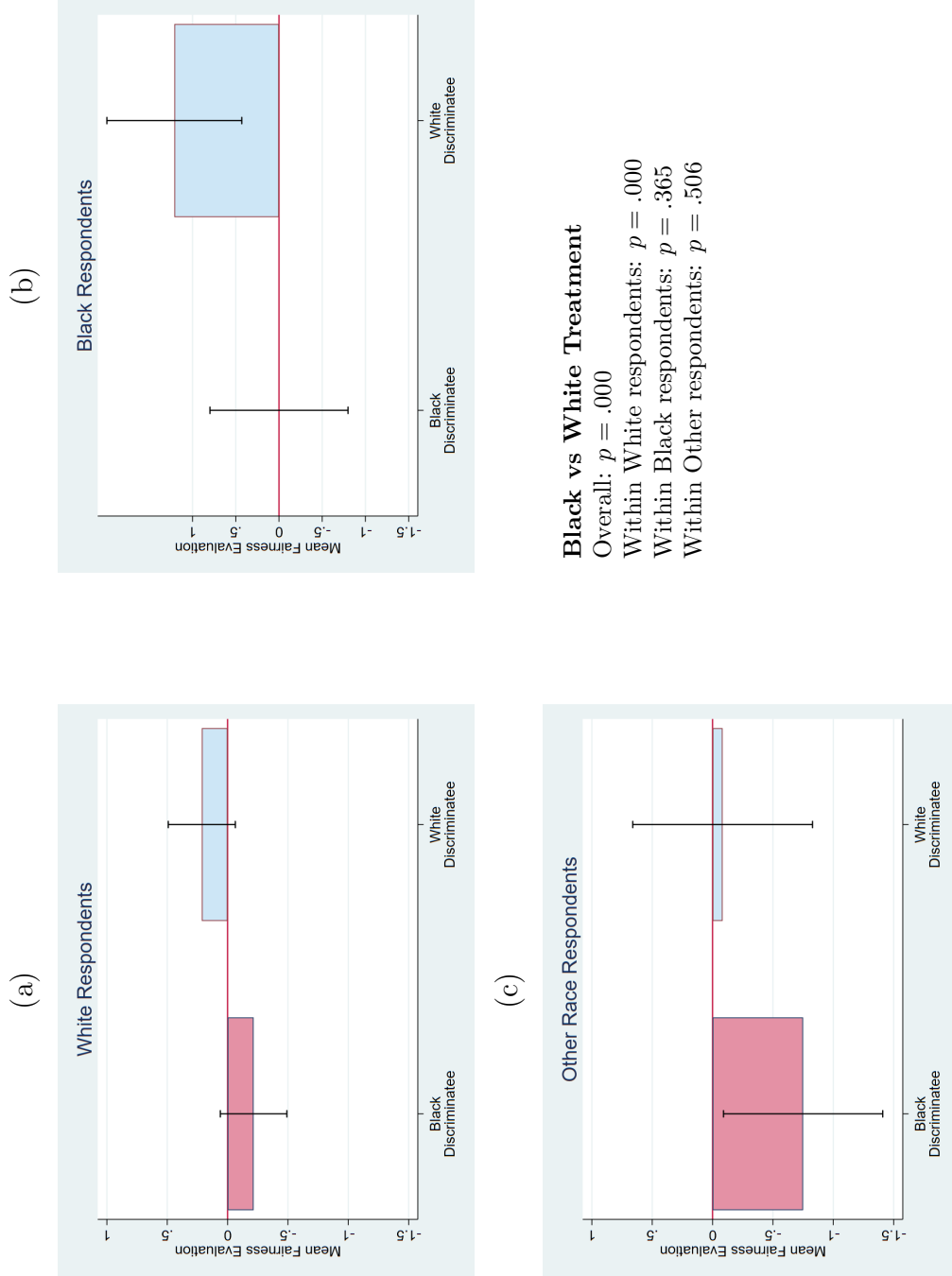
Less vs More Justifiable Treatment

Overall: $p = .000$
 Within Less-Justifiable: $p = .000$
 Within More-Justifiable: $p = .000$

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

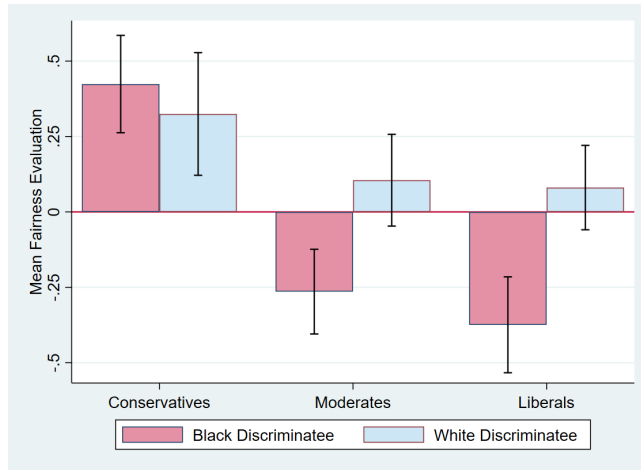
Within Black Discriminatees, less-justifiable scenarios are 0.898 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.953 units less fair. A test for equality of the Less versus More Justifiability Gap between the Black and White treatment yields $p = .679$.

Figure 2.4: Fairness Ratings by Respondent Race and Discriminatee Race



Notes: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. All p -values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) across all three racial groups yields $p = .739$.

Figure 2.5: Fairness Ratings by Political Orientation and Discriminatee Race

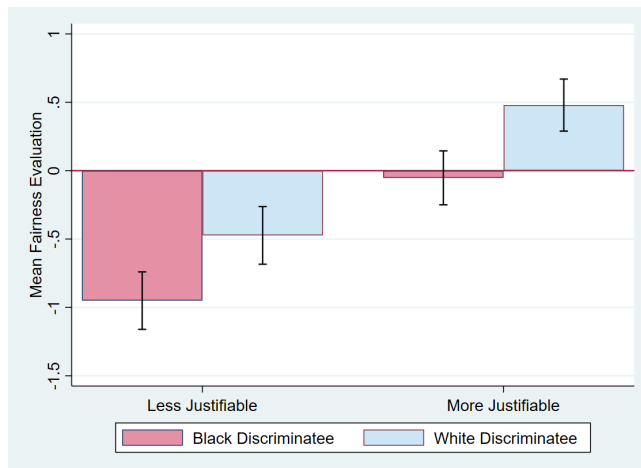


Black vs White Treatment

Overall: $p = .000$
 Within Conservatives: $p = .365$
 Within Moderates: $p = .000$
 Within Liberals: $p = .000$

Notes: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. All p -values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields $p = .567$. A test for equality between conservatives and (moderates + liberals) yields $p = .001$.

Figure 2.6: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent’s Political Leaning

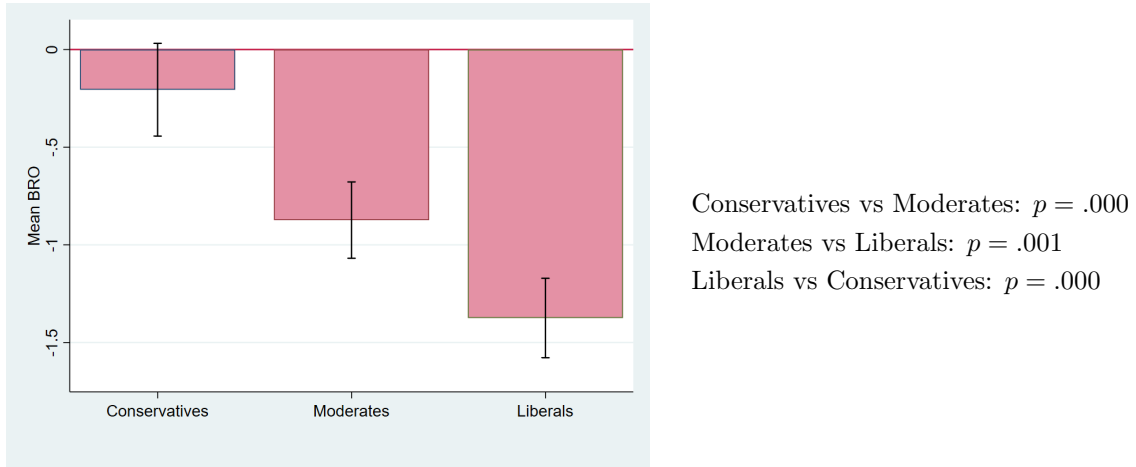


Black vs White Treatment

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

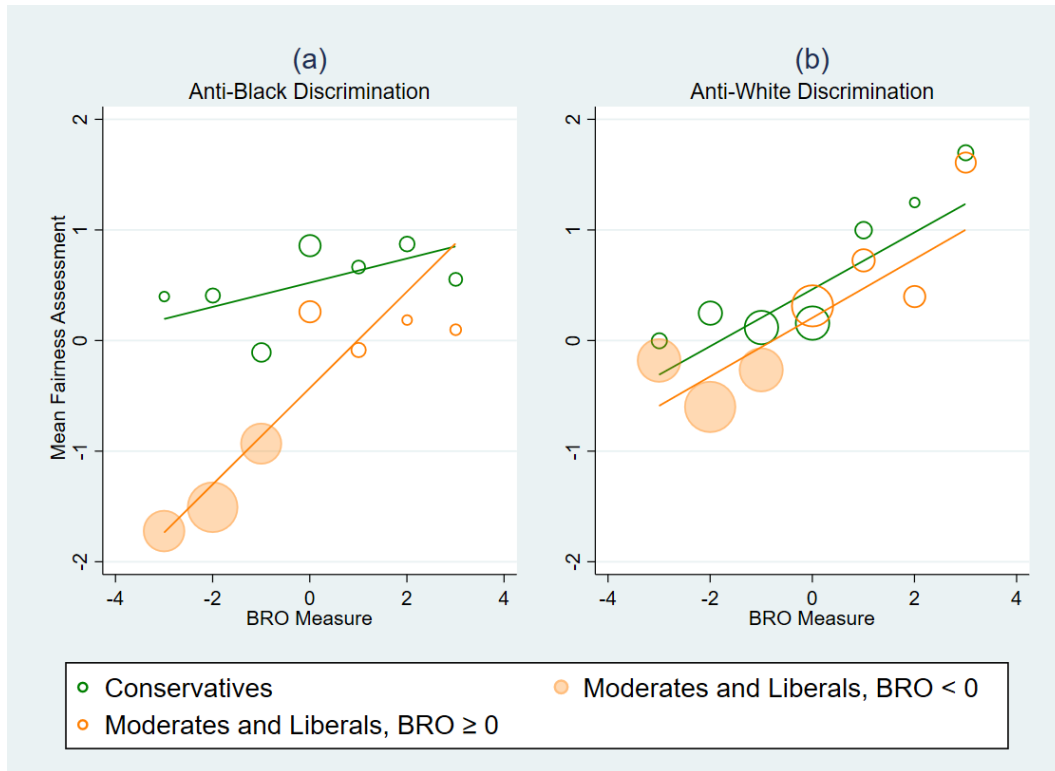
Notes: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. All p -values are clustered by respondent. A test for equality of the Less versus More Justifiability Gap across Conservatives, Moderates, and Liberals yields $p = .590$.

Figure 2.7: 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). Figure is based on Stage 1 observations only. 95% confidence intervals are shown. All p -values are clustered by respondent. A test for equality of BRO across all three political groups yields $p = .577$.

Figure 2.8: 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.

- 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 ($p = .000$)

Chapter 3

Do Perkins Loans Come With Perks?

3.1 Introduction

The federal government has been playing a significant role in maximizing access to higher education since it implemented the Higher Education Act of 1965 by allocating funds towards need-based financial aid programs. A vast economic literature has assessed the efficacy and consequences of these programs on the outcomes of student recipients. For example, many studies have estimated the effects of various programs on college access and attainment (Seftor & Turner, 2002; Dynarski, 2003; Bettinger, 2004; Turner & Marx, 2018). Other studies focus on the implications of these programs on the post-graduation returns for students (Denning et. al, 2019; Gervais & Ziebarth, 2019). While many of these studies have focused on prominent programs, such as Pell grants and Stafford loans, little attention has been paid to the effects of Perkins loans, which includes a unique *perk* that rewards recipients who enter certain career paths.

This study estimates the effects of the Perkins Loan Program on undergraduate students' non-academic work status and hours (i.e., labor supply) and choice of STEM college major and post-graduation occupation. This program provided low-interest loans

to students with relatively high levels of financial need. This program allowed the loans to be forgiven if recipients opted to work in certain public service occupations (e.g., a K-12 teacher). In other words, the loan was converted to a non-repayable grant, conditional on the recipient's occupation. Although the Perkins Loan are no longer issued as of 2018, its analysis could highlight the costs and benefits of either revitalizing it or introducing a similar program.¹

This study contributes to an emerging body of literature that considers the effects of financial aid on student labor supply and academic performance. This literature is motivated by increasing trends in students delaying their completion of 4-year degrees and taking up non-academic-related jobs amid rising college costs (Bound et al., 2010; Scott-Clayton, 2012). The primary idea behind this area of study is that any form of financial burden or debt is detrimental to enrolled college students by giving them the incentive to spend more time working a non-academic job than studying. Consequently, higher work hours may lead to lower academic performance (Stinebrickner and Stinebrickner, 2003; Kalenkoski and Pabilonia, 2010). Several studies focus on the effects of need-based grant programs, such as Pell grants, and find some evidence that they reduce students' work hours (Broton et al., 2016; Denning, 2019; Kofoed, 2022). Kofoed (2022) also finds evidence that such forms of financial aid increase students' grades by roughly a third of a letter grade, on average.

Unlike grant programs, loans need to be repaid and tend to have relatively muted effects on student outcomes (Heller, 2008). For example, Dynarski (2004) finds little evidence that student loan expansions in the U.S. in the early 1990s led to an increase in college attendance. Furthermore, Park and Scott-Clayton (2018) finds that much of the reduction in student labor supply resulting from federal financial aid can be attributed

¹Federal Perkins Loans Program was phased out in 2015 due to federal budgetary issues but proliferated until 2017. However, Congress failed to renew the program and rendered it expired on September 30, 2017. Final loan disbursements were allowed through June 30, 2018.

to the effects of Pell Grants rather than loans. Since Perkins loans can be forgiven, they may have similar effects on student outcomes as grant programs. Furthermore, Perkins loans could encourage recipients to persist throughout their college education and motivate them to choose a relevant career path dictated by the program's guidelines for loan forgiveness.

Hence, this study also contributes to the literature by investigating the impacts of a financial aid program that grants students conditional loan forgiveness. It determines whether the Perkins loans incentivizes students to choose a career path that could forgive their loan amount and therefore offers similar perks as grant programs, such as reduced work hours and improvement in academic performance. In doing so, this study illustrates whether this program was effective at delivering its intended incentives to students receiving the loans.

This study uses a regression and matching approach to estimate the average treatment effect on the treated (ATT) for the Perkins Loan program. It ultimately finds that the program decreased the likelihood of students working a non-academic job and their work hours. However, it also finds that the program led to a marginally small reduction in students' grades. This pattern could reflect Perkins loan recipients taking up more rigorous coursework or reallocating their time toward leisure or career-academic-related extracurricular activities. Finally, this study finds that the program had any effect on students' choice of pursuing a post-graduation occupation that could qualify them for loan forgiveness.

This paper proceeds as follows. Section 2.3 provides some institutional background and details behind the Perkins Loan Program. Section 3.3 describes the data sets used in all analyses. Section 3.4 discusses the study's empirical strategies. Section 3.5 discusses the results. Finally, Section 3.6 concludes.

3.2 Background of the Perkins Loan Program

The Federal Perkins Loan Program, named for the late Kentucky Representative Carl D. Perkins, provided undergraduate and graduate students with a need-based loan. These loans were dispersed by the U.S. Department of Education. They had a fixed 5% interest rate on a 10-year repayment period, and they accrued nine months after the borrower ceased to be a student. Undergraduate borrowers were limited to \$5,500 per year, with a cumulative limit of \$27,500. Graduate students were limited to \$8,000 per year with a total limit of \$60,000, including any undergraduate Perkins loan debt.

As discussed in Section 3.1, this aid program had a unique feature: recipients could get their Perkins loan amount fully canceled if they opted to work in certain public service occupations. These include being an elementary/secondary school teacher, nurse or medical technician, speech pathologist, military service, public defender, faculty member at a Tribal College or University, firefighter, child or family services provider, or law enforcement officer.² Loan cancellation, however, occurred in increments across a 5-year span. Specifically, 15% of the original debt was forgiven for the first year of service, 15% for the second year, 20% for the third year, 20% for the fourth year, and 30% for the fifth year.

To be eligible to apply for a Perkins loan, a student must be enrolled at a Title IV participating institution as an undergraduate, graduate, or professional student.³ Furthermore, they must have submitted a free application for federal student aid (FAFSA).

²Recipients who chose to work as a librarian and speech pathologist must obtain a Master's degree in their respective field and serve in an eligible elementary or secondary school to qualify for loan cancellation.

³A Title IV institution is a college or university that processes U.S. federal student aid. Therefore, qualifying students at these colleges can receive grants or loans subsidized by the Department of Education. These institutions include public, private non-profit, and proprietary schools. To qualify as a Title IV institution, a school must abide by certain regulations, e.g., be accredited or pre-accredited by an accrediting agency recognized by the Education Department to grant accreditation or pre-accreditation status.

And to do so, they must be either a U.S. citizen or qualifying non-citizen. The information provided by a student's FAFSA is sent to the institution they attend, whose financial aid office uses it to determine the student's cost of attendance (COA) and expected family contribution (EFC). A student's COA is the sum of tuition, fees, and additional educational expenses (e.g., books, room & board, etc.). Financial aid offices consider factors when determining COA, such as the student's attendance status, dependency, and type of housing. The EFC is the student's estimated ability to pay for their education expenses.

Once an institution's financial aid office calculates a student's EFC and COA, they use these measures to determine the student's eligibility for federal, state, and institutional grants as well as the amount they receive from each program. Students will qualify for some need-based grant program if their COA exceeds their EFC (i.e., $COA - EFC > 0$), and the award amounts will often be determined by the gap size between these two measures.⁴ However, if a student's COA exceeds their EFC even after accounting for grant amounts they received, then that student can qualify for federal loans to cover their remaining educational expenses (Gervais & Ziebarth, 2019). Hence, a student's qualification for some loan program is determined by their excess financial *need*, or $need_i$, which is defined by the following equation

$$need_i = COA_i - EFC_i - G_i \quad (3.1)$$

where G_i is the sum of grant amounts received from government or institutional sources. In other words, a student (i) qualifies for some federal loan amount if $need_i > 0$.

Unlike some federal loan programs, such as Stafford Loans, the Perkins Loan Program does enforce definitive rules for determining qualification and loan amounts to be employed uniformly among all Title IV colleges and universities. Student qualification

⁴An example of such program is the Federal Pell Grant program. A detailed description of how award amounts are determined for this program can be found in Kofoed (2022).

for this program is left to the discretion of each institution’s financial aid office, although the Department of Education instructs them to use $need_i$ in determining Perkins loan amounts.⁵ Therefore, any “cutoffs” or “thresholds” in $need_i$ for qualifying for Perkins loans (i.e., being assigned “treatment”) are unobservable and likely to be heterogeneous across institutions. As discussed in Section 3.4, the lack of this information limits the alternative strategies for analyzing this program and identifying its effects, especially with regards to addressing endogeneity.

3.3 Data

This study uses data from the *National Postsecondary Aid Survey* (NPSAS), a nationally-representative cross-sectional survey that studies the characteristics of students at U.S. colleges & universities and how they finance their education. Each wave of this survey is administered every four years by the Institution of Education Sciences’s (IES) National Center of Education Statistics (NCES), a branch of the U.S. Department of Education. The NPSAS provides a rich set of individual-specific variables such as students’ financial aid amounts (e.g., grants and loans) from various sources, FAFSA information (e.g., EFC), enrollment status, GPA, and demographic characteristics. It also contains information on each students’ institution of attendance, such as its sector (e.g., 4-year private non-profit) and degree of admissions selectivity.

This study also utilizes data from *Baccalaureate & Beyond* (B & B), a survey that is administered by the NCES. The B & B follows up on the activities (e.g., occupational choice) on select respondents from every other NPSAS survey wave shortly after they

⁵The outline of the Department of Education’s guidelines for selecting students for the Perkins Loan Program can be found in Title 34, §674.10 of the Code of Federal Regulations (CFR). It establishes that this program should be prioritized to students with *exceptional* financial need. However, it specifically states, “The institution shall define exceptional financial need for the purpose of the priority [awarding Perkins loan amounts]...and shall develop procedures for implementing that priority.”

graduate (i.e., at least 1 year after). Therefore, only a fraction of NPSAS respondents is chosen to participate in this follow-up survey. However, B & B is also designed to be nationally-representative. The availability of student identifiers within both the NPSAS and B & B allows for data linkage between these two surveys. Linkage between these two surveys enables the observation of Perkins loan recipients' and non-recipients' exact career paths following graduation.⁶

The specific outcome variables utilized by this study include the weekly non-academic work hours of students, grade point average (GPA), STEM major status, and an indicator for pursuing an occupational that could grant a student loan forgiveness.⁷ GPA is standardized to improve the comparability of grades between institutions. This study also utilizes control variables, including financial need, a set of indicators for the selectivity of a student's institution (e.g., moderately selective, minimally selective, etc.), and a set of indicators for underrepresented minority background status, gender, full-time status, seeking a Bachelor's degree, residing in the same state as the institution (i.e., state residency), and attending a private institution.⁸ Financial need is constructed in accordance to its definition described in the previous section, i.e., it is a student's COA minus EFC minus the sum of government and institutional grants they received. This measure is adjusted for inflation to 2016 U.S. dollars (USD) and is expressed in thousands.

Specifically, this study pools together the 2003-2004, 2007-2008, 2011-2012, and 2015-2016 waves of the NPSAS. The 2007-2008 and 2015-2016 waves include student respon-

⁶The B & B survey records respondents' occupation with a six-digit SOC code based on the Bureau of Labor Statistics' definitions. The high level of detail of these codes enable the identification of the very occupations that could qualify students for Perkins loan forgiveness.

⁷This study excludes work hours from Federal Work Study. This need-based program provides students funds via a part-time job with participating employers (e.g., a non-profit organization, government agency, etc.) to cover their college costs. The government dictates how much a student can work under this program, so students would have less flexibility toward working a job off campus versus work study.

⁸Students are considered dependent if they are over the age of 24, has dependents, is married, or is a military veteran. Otherwise, they are considered independent. Also, the choice of STEM major status corresponds to the fact that most of the loan-forgiveness granting occupations are unrelated to STEM fields, with the exception of medical technicians and nurses.

dents who were chosen for the B & B follow-up surveys (i.e., the 2008-2009 and 2016-2017 B & B waves). The selection of these waves is based upon the availability of some key variables, in particular, the work hours of enrolled students.⁹ This study includes undergraduate students from 4-year public and private not-for-profit colleges & universities since a significant fraction of Perkins loan recipients (i.e., about 78.7% of them) attended these types of institutions. Also, this study includes students who applied for federal financial aid (i.e., submitted a FAFSA form) and have financial need (i.e., $need_i > 0$). Finally, it omits students who attended multiple institutions in a single year and those from institutions that did not disperse Perkins loans.

The restrictions described above results in a preliminary sample size of approximately 54,080.¹⁰ Of this sample, about 11,040 were selected to participate in the B & B follow-up surveys. Perkins loan recipients comprise of approximately 14.2% and 13.6% from the NPSAS and B & B, respectively. The summary statistics for all variables can be found in the left panel of Table 1.1. It shows that the Perkins loan recipients and non-recipients are unbalanced in almost all of these variables. However, some of these differences are small in magnitude. Table C1.1 of Appendix C contains the summary statistics for students included in the B & B, and similarly, the recipient and non-recipient groups are mostly unbalanced.

Table 3.2 compare the shares of Perkins loan recipients and non-recipients across 12 categories of college majors. The former indicates that a relatively greater share of Perkins loan recipients enters some STEM-related majors (e.g., life and physical sciences) relative to non-recipients, except for computer/information science.¹¹ A relatively higher

⁹Earlier NPSAS survey waves do not record students' work hours.

¹⁰In accordance to the IES's data security requests, all (unweighted) sample sizes reported throughout this paper are rounded to the nearest tens. This measure helps ensure the disclosure protection (i.e., privacy) of respondents surveyed for the NPSAS and B & B.

¹¹The difference in shares between recipients and non-recipients for math is insignificant (i.e., $p = .0.175$).

fraction of non-recipients tend to be either undeclared or not enrolled in a degree program. Finally, a relatively greater fraction of non-recipients enter business/management, other technical/professional fields, education, and health-related fields.

Table 3.3 shows the shares of Perkins recipients versus non-recipients among various loan-forgiveness-granting occupations. It indicates that the overwhelming majority of recipients chose not to enter a profession that qualifies for loan forgiveness. Of the recipients choosing a loan-forgiveness-granting career path, the majority opted for a primary/secondary teaching route.¹² Finally, these tables show that the shares for non-recipients across majors and occupations are similar to those of recipients. And the differences in shares between these two groups of students are insignificant.

3.4 Empirical Strategies

Specifically, this study estimates the average treatment effect on the treated (ATT) for the Perkins Loan program on all outcomes of interest. The estimation of this treatment effect falls within the *potential outcome* framework where some outcome for student i is defined as:

$$Y_i = \begin{cases} Y_i(1) & , \text{ if } P_i = 1 \\ Y_i(0) & , \text{ if } P_i = 0 \end{cases} . \quad (3.2)$$

Above, P_i is an indicator for being a Perkins loan recipient (i.e., receiving treatment) and $Y_i(1)$ and $Y_i(0)$ are the potential outcomes for being a recipient and non-recipient,

¹²None of the Perkins loan recipients from the B & B sample appears to have entered a career in firefighting or related occupations. Furthermore, none seem to have enrolled in a graduate program that would enable them to pursue a career in the library sciences or speech pathology.

respectively. Under this framework, the estimand of interest is defined by the following:

$$\tau^{att} = \mathbb{E}(Y_i(1) - Y_i(0)|P_i = 1). \quad (3.3)$$

The above equation is often referred to as the *population* ATT.

An advantage of estimating ATT over another parameter of interest, the average treatment effect (ATE), is that the identification assumptions for the former are relatively less stringent than those required to estimate the latter, which are less likely to be plausible in practice.¹³ Specifically, the identification of the ATT hinges upon the following assumptions: (i) *unconfoundedness for the untreated* and (ii) *partial overlap* (Heckman, Ichimura, and Todd, 1998; Imbens, 2004).¹⁴ The first assumption posits that treatment status is as good as random among the potential outcomes for non-recipients, conditional on observable factors. The second posits that no student can receive Perkins loans with certainty, regardless of their observables.

This study explores two possible strategies for estimating the ATT, which are described in the following sub-sections: ordinary least squares and (nearest-neighbor) matching.

3.4.1 OLS

A simple, preliminary approach to estimating the ATT for the Perkins loan program on all outcomes of interest is ordinary least squares (OLS). Essentially, this strategy

¹³The identification of the ATE relies on (i) (*strong*) *unconfoundedness* and (ii) *overlap*. The former posits that $Y_i(1), Y_i(0) \perp P_i | \mathbf{X}_i$ for all i , where \mathbf{X}_i is a vector of covariates. The latter requires $0 < \mathbb{P}[P_i = 1 | \mathbf{X}_i] < 1$ for any $\mathbf{X}_i \in \mathcal{X}$.

¹⁴Formally, *unconfoundedness for the untreated* says for each student i , $Y_i(0) \perp P_i | \mathbf{X}_i$ where \mathbf{X} is a vector of (observable) covariates. And *partial overlap* states $\mathbb{P}[P = 1 | \mathbf{X}] < 1$ for any $\mathbf{X} \in \mathcal{X}$ where \mathcal{X} is the support of \mathbf{X} .

entails estimating the following regression specification:

$$Y_i = \alpha_s + \tau^{ols} P_i + \mathbf{X}_i' \boldsymbol{\theta} + \epsilon_{is} \quad (3.4)$$

Similar to the notation displayed earlier, Y_i is the outcome variable, P_i is an indicator for receiving the Perkins loan (i.e., treatment), and \mathbf{X}_i is a vector of covariates. Specifically, \mathbf{X}_i includes financial need and sets of indicator variables for coming from a URM background, gender, full-time status, dependency, seeking a BA, being a resident within the same state as the college/university of attendance, attending a private institution, and the selectivity of the students' institution. Finally, α_s denotes the survey wave fixed effects.

A significant challenge of employing OLS to study Perkins loans is choosing an adequate non-recipient comparison group for recipients from the sample. One way to address control group selection is to restrict the sample across different thresholds for financial need (e.g., all students with $need_i > \$5,000$) when estimating equation (3.2) since that measure plays a significant role in selecting participants for the program. Nevertheless, OLS estimates for the ATT within these analyses may be subject to selection bias since Perkins loan take-up is endogenous, i.e., students may choose to turn down an offer for these loans. This endogeneity issue is motivated in Tables C2.1 and C2.2 of Appendix B, which show that take-up depends on various characteristics of students, such as their financial need, selectivity of their institution, and survey wave year. Furthermore, these patterns are also evident across varying thresholds of students' financial need.

3.4.2 Matching

Simple Matching

The options for potential empirical strategies that could mitigate the issues described above are limited since the NPSAS and B & B data sets do not indicate whether a student was offered Perkins loans. Furthermore, as discussed in Section 3.2, thresholds for $need_i$ to qualify for the program are unobservable and likely heterogeneous across colleges and universities.¹⁵ Therefore, this study applies a nearest-neighbor matching (NNM) approach to exploit $need_i$ as the primary determinant for receiving a Perkins loan – i.e., pair recipients with counterfactual non-recipients with close-proximity levels of financial need. And in particular, NNM provides the benefit of generating a balanced sample when estimating the ATT for the Perkins Loan program.

The ATT is estimated as follows. First, every loan recipient i with some level of financial need ($need_i$) and a set of discrete characteristics ($\widetilde{\mathbf{X}}_i$) is matched with a non-recipient, $j(i)$, with the same $\widetilde{\mathbf{X}}_i$ (e.g., both attend a private institution and are full-time students) and level of financial need in “close proximity” to i ’s level. Formally, the matching non-recipient $j(i)$ must satisfies the following criteria:

$$\|need_{j(i)} - need_i\| = \min_{l:P_l=0} \|need_l - need_i\| \quad (3.5)$$

$$\widetilde{\mathbf{X}}_i = \widetilde{\mathbf{X}}_{j(i)} \quad (3.6)$$

where $\|\cdot\|$ denotes the Euclidean metric.¹⁶ Recipient i may be matched with more than one non-recipients when they are “tied” in observables. I.e., there may be $M \geq 1$ number

¹⁵Some studies, such as Turner and Marx (2018), are able to exploit cutoffs in financial need to identify the local average effects of certain federal programs via fuzzy regression kink designs.

¹⁶NNM often employs the Mahalanobis metric, which essentially normalizes the Euclidean norm by the sample variance-covariance matrix. This metric is best utilized when units are matched by two or more continuous covariates to ensure each are scaled comparably when measuring distance. However, with one continuous covariate, $need_i$, the choice of the Euclidean versus Mahalanobis metric is inconsequential.

of non-recipients that satisfy (3.4) and (3.5). Furthermore, matching is conducted *with replacement*. Therefore, non-recipient $j(i)$ may serve as a match for more than one recipients.

Second, the outcome for non-recipient $j(i)$, Y_j , is imputed as a proxy for the counterfactual outcome for recipient i (i.e., their outcome had they not received a Perkins loan). This counterfactual is defined as:

$$\hat{Y}_i(0) = \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j \quad (3.7)$$

where $\mathcal{J}_M(i) = \{j_1(i), j_2(i), \dots, j_M(i)\}$ is the set of indices of the first M matches for student i . As discussed, multiple matches (i.e., $M > 1$) results from ties in observables among non-recipients in $\mathcal{J}_M(i)$. In these cases, as shown by equation (3.7), the outcomes across these matches (i.e., the Y_j 's) are averaged across M . Finally, the ATT is estimated using the following estimator:

$$\hat{\tau}^{match} = \frac{1}{n_1} \sum_{i=1}^n P_i \left(Y_i - \hat{Y}_i(0) \right) \quad (3.8)$$

where n_1 is the number of recipients in the sample. Inference for the point estimates resulting from this estimator is based on robust standard errors derived by Abadie and Imbens (2006).

The setup of the NNM algorithm and the composition of treated students in the sample sheds light on another critical advantage of estimating the ATT over the ATE. In particular, the relatively large size of the non-recipient group helps ensure the availability of close matches for recipients. And as shown in Table 1.1, a significant share of students in the sample are non-recipients. On the other hand, ATE estimation relies on a *symmetric* matching algorithm that pairs non-recipients with counterfactual recipients

satisfying (3.5) and (3.6). Given that the share of recipients within the sample is small (i.e., less than 20%), matching discrepancies are likely to arise from this algorithm which could ultimately hinder the balance in financial need within the matched sample and introduce bias.

Bias-Corrected Matching

A caveat of the matching estimator from equation (3.8) is that it may be biased due to matching discrepancies (Abadie and Imbens, 2006). This issue often arises when matching is conducted on *two or more* continuous variables. The bias takes the following form:

$$B_i = \mu^0(\mathbf{X}_i) - \mu^0(\mathbf{X}_{j(i)}). \quad (3.9)$$

where \mathbf{X}_i is a vector of covariates, $\mu^0(\mathbf{X}_i) = \mathbb{E}(Y_i | \mathbf{X}_i = \mathbf{x}_i)$ and $\mu^0(\mathbf{X}_{j(i)}) = \mathbb{E}(Y_i | \mathbf{X}_{j(i)} = \mathbf{x}_{j(i)})$. Within the context of this study, \mathbf{X}_i solely contains $need_i$. The bias may persist within a large sample, i.e., matching estimators may not be \sqrt{n} -consistent and its order is independent of the number of discrete covariates including within the matching procedure (Imbens and Wooldridge, 2009).

To address this issue, Abadie and Imbens (2011) proposes a bias-corrected estimator that subtracts an estimate of B_i from equation (3.7). This estimator takes the form of

$$\hat{\tau}^{match'} = \frac{1}{n_1} \sum_{i=1}^n P_i [(Y_i - \hat{Y}_i(0)) - (\hat{\mu}^0(\mathbf{X}_i) - \hat{\mu}^0(\mathbf{X}_{j(i)}))]. \quad (3.10)$$

The $\hat{\mu}^0$'s correspond to the predicted/fitted values obtained from regressing Y_i on \mathbf{X}_i (or, $need_i$).¹⁷ Therefore, $\hat{\mu}(\mathbf{X}_i)$ corresponds to the predicted value for recipient i while

¹⁷For example, if one estimates the following regression, $Y_i = \gamma_0 + \gamma_1 \mathbf{X}_i + \eta_i$, then the fitted values serve as the estimates of μ^0 . I.e., $\hat{\mu}^0(\mathbf{X}_i) = \hat{\gamma}_0 + \hat{\gamma}_1 \mathbf{X}_i$.

$\hat{\mu}(\mathbf{X}_{j(i)})$ is that of matching non-recipient $j(i)$.

3.5 Results

3.5.1 Results for OLS

The results of estimating equation (3.2) are provided in Columns 1-4 of Table 3.4. Each column corresponds to sample restrictions imposed on students' financial need amounts (e.g., $need_i > \$5,000$). These estimates are primarily consistent across these various restrictions on financial need. They suggest that the program led to a decrease in the likelihood of working, work hours, and GPA. However, they also indicate that the program has little effect on students' chances of taking up a STEM major in college and pursuing a career path that would qualify them for loan forgiveness. The point estimates for the latter outcome, however, are relatively noisy. As discussed in Section 3.1, these results may be subject to endogeneity, so the following sections re-estimate the effects of Perkins loans via matching.

3.5.2 Results for Matching

The matching algorithm imposes no sample restrictions on students' financial need (i.e., it includes anyone with $need_i > 0$) and matches each of them with at least 1 student, with the maximum number of matches per recipient being 10 due to ties in distances.¹⁸ And the B & B sub-sample has between 1 and 3 matches per recipient. However, this algorithm led to about 12.9% of observations needing to be discarded due to unmatched

¹⁸One advantage of having multiple matches per student recipient is that this reduces the sampling variance of the matching estimator. But multiple matches come with a trade-off: this introduces bias by increasing the average covariate discrepancy within pairs (Imbens and Rubin; 2016). However, the overwhelming majority of Perkins loan recipients (i.e., about 7,080 out of 7,260) received at most one match. And less than 20 recipients received over three matches.

students. The sizable number of non-matches comes from the fact that these students are *exactly* matched on many discrete covariates, such as gender and survey wave. Many of these discarded observations are loan non-recipients. The right panel of Table 3.1 contains the summary statistics for this trimmed, *raw* sample used for matching. Although the share of Perkins loan recipients is slightly higher within this sample (i.e., 15.4%), much of these summary statistics are similar to those of the “full” sample of the left-hand panel. The relative stability of these summary statistics could be owed to the already large size and representativeness of the NPSAS data set.

Similarly, 17.8% units from the B & B sub-sample are also discarded due to non-matches. The summary statistics of the trimmed, *raw* B & B sub-sample can be found in the right panel of Table C1.1, and these are mostly comparable to those of the full B & B sub-sample. However, this trimmed sub-sample contains noticeably fewer shares of URM students in both the Perkins loan recipient and non-recipient groups.

The NNM algorithm generates a matched sample containing about 14,520 observations – i.e., each of the 7,260 recipients receives a matching non-recipient for estimating the ATT. Table 3.5, which displays the summary statistics for the resulting matched sample among all covariates between recipients and non-recipients, indicates that NNM improves covariate balance. As expected, the discrete covariates (e.g., URM status) are perfectly balanced. The difference in financial need between non-recipients is smaller than that of the right panel of Table 3.1 and insignificant ($p = .520$). And, as shown in Figure 3.1, the distributions for financial need for recipients and non-recipients are very similar and are indistinguishable ($p = .975$). As shown in Appendix C.1, the matching procedure is conducted separately for the B & B sub-sample and indicates that NNM improves balance within that sub-sample. Finally, the distance between matched pairs of students is relatively small at 0.0112 (i.e., about \$11.2).

The point estimates from equation (3.8) for each outcome can be found in Column 5

of Table 3.4. Most of the point estimates are comparable to the OLS ones. In particular, it indicates that Perkins loan recipients tend to work less for a non-academic job, i.e., on average, they are 0.0569 points less likely to work while enrolled and spend about 1.563 fewer hours toward working. Furthermore, on average, loan recipients' GPAs are 0.116 standard deviations lower than non-recipients. Moreover, Table 5 shows that the program increases the likelihood of majoring in a STEM field by a small amount, i.e., 0.00378 points. Finally, the program has little effect on the possibility of pursuing a career path that could grant loan forgiveness, although this point estimate has a wide confidence interval.¹⁹

Column 6 of Table 3.4 replicates the analyses from Column 5 using the bias-corrected estimator from equation (3.9). All of these estimates are similar to those of the latter column. The minimal bias resulting from any matching discrepancies could be attributed to students being matched on only one continuous variable, financial need.

3.6 Discussion

This study estimates the effects of the Perkins Loan Program on the work status and hours of enrolled college students, as well as their choice of STEM major and occupation. It finds that the program may not have all the perks offered by other financial aid forms. While it reduced students' work hours, it also hurt their grades. Furthermore, it finds limited evidence that Perkins loans increased students' likelihood of majoring in a STEM field. But there is no evidence that recipients are likely to enter an occupation that could qualify them for loan forgiveness.

The program's negative impact on academic performance could be attributed to loan

¹⁹The results described in this paragraph are robust to a caliper of 0.1. Here, matches must be within \$100 of each other in financial need, and this mitigates outliers in wide matched distances. Furthermore, these results are robust to a sample with financial need greater than \$10,000.

recipients being more likely to take a STEM major or complete related courses. Indeed, Arcidiacono et al. (2012) notes that science, engineering, and economics courses tend to give out lower grades than humanities and social science courses. However, the matching point estimate for the choice of STEM major is small and noisy. Also, the effect size for academic performance is relatively small, i.e., a 0.0116 decrease in standard deviations translates to a decline in a hundredth of a grade point, assuming a 4.0 scale. Alternatively, recipients may have shifted their time from work toward leisure or academic-related extracurricular activities. However, the NPSAS does not provide information on the amount of time students spend on such activities.

The lack of evidence that students choose occupations granting loan forgiveness could be attributed to these alternatives yielding relatively lower returns. And as discussed in Section 3, the overwhelming majority of Perkins loan recipients do not take up these occupations. However, the Bureau of Labor Statistics (BLS) data suggests the average annual earnings of these options (i.e., roughly \$58,500) is not much higher than the overall mean earnings of \$58,260 among all occupations.²⁰ Alternatively, recipients may have lacked knowledge of this program's forgiveness perk. Indeed, federal law does not require colleges and universities to provide clear, standardized information on all their financial aid offers (U.S. Government Accountability Office, 2023). Thus, they may not follow the best practices in providing transparent information regarding these offers and, therefore, facilitate information barriers that hinder recipients from understanding the full benefits they could receive from Perkins loans.

This study has a few caveats. First, a matching approach does not entirely resolve the endogeneity of Perkins loan take-up.²¹ However, this empirical strategy is an optimal

²⁰The BLS keeps track of estimated earnings across various occupations. These observations are based on observations from 2021.

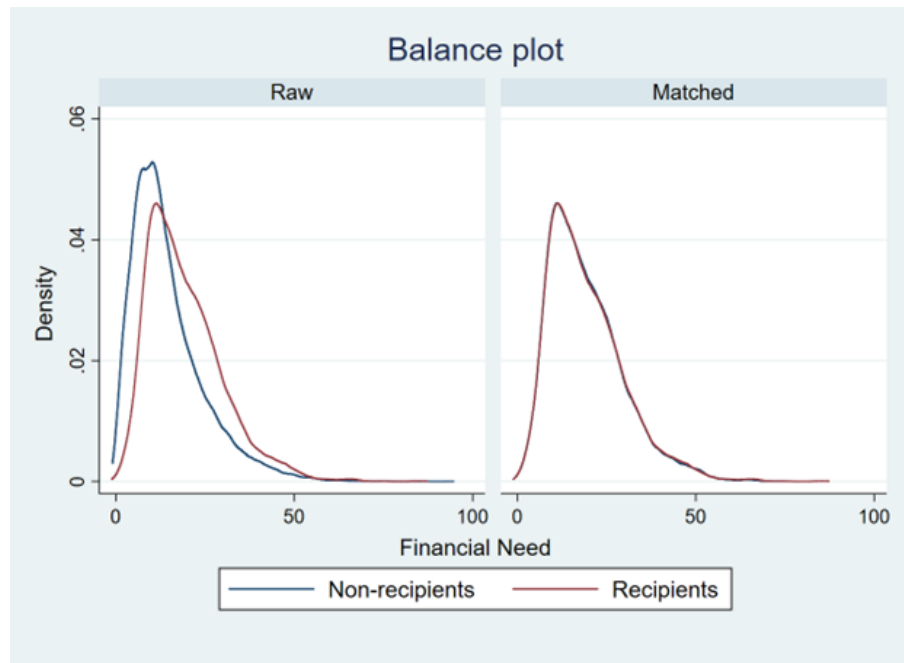
²¹As discussed by Heckman, Ichimura, and Todd (1997), the framework behind matching does not incorporate separability between observables and unobservables.

alternative given the scope of the NPSAS and B & B data sets. Second, there are increasingly complex interactions between different forms of financial aid (e.g., Pell grants and institutional grants), which makes it inevitably difficult to fully isolate the effects of one program (Turner, 2014). Thus, the estimates displayed in Table 3.4 are likely to illustrate *descriptive*, rather than causal, relationships. Finally, while the B & B revisits respondents across a few waves beyond graduation, later follow-up survey data was unavailable. This issue hinders this study from observing the long-run occupational choices of students (i.e., up to 10 years after graduation).

In short, the Perkins Loan Program may not have delivered the intended career incentives to its recipients. However, future studies could incorporate later B & B follow-up waves to estimate the effects of Perkins loans on occupational choices in the long run. Also, future studies could further investigate the impact of financial aid programs that offer conditional loan forgiveness. For example, other forms of financial aid, such as Texas's B-On-Time loan, forgives recipients' loan amounts upon successful degree completion from their college/university of attendance. Such considerations for conditional loan-forgiving aid could help determine whether they deliver the right incentives that affect key student outcomes, including academic performance and timely graduation.

3.7 Figures

Figure 3.1: Distribution for Financial Need



Notes: The left panel of this figure compares the distribution for financial need between Perkins loan recipients and non-recipients for the raw, unmatched sample. The right panel displays that of the matched sample. This figure serves as a check for how well matching improves the balance of covariates, in particular, financial need. Here, the distributions between recipients and non-recipients in the matched sample are statistically indistinguishable per a Kolmogorov-Smirnov test ($p = .975$).

SOURCE: 2003-2004, 2007-2008, 2011-2012, and 2015-2016 NPSAS.

3.8 Tables

Table 3.1: Summary Statistics – Unmatched Sample

	Full Sample			Sample for Matching		
	Recipients (1)	Non- recipients (2)	diff. <i>p</i> -val. (3)	Recipients (4)	Non- recipients (5)	diff. <i>p</i> -val. (6)
Worked at all	0.52 (0.50)	0.62 (0.49)	0.00	0.52 (0.50)	0.61 (0.49)	0.00
Hours worked while enrolled	10.68 (13.38)	14.38 (15.06)	0.00	10.49 (13.32)	13.78 (14.62)	0.00
Standardized GPA	-0.11 (1.00)	0.00 (0.99)	0.00	-0.10 (1.00)	0.02 (0.98)	0.00
STEM Major	0.38 (0.49)	0.36 (0.48)	0.01	0.39 (0.49)	0.38 (0.48)	0.24
Loan-forgiveness Occupation	0.15 (0.36)	0.16 (0.37)	0.29	0.15 (0.36)	0.17 (0.37)	0.41
Financial need	19.37 (10.48)	14.43 (10.13)	0.00	19.34 (10.49)	14.43 (10.08)	0.00
Underrepresented Minority	0.25 (0.44)	0.24 (0.43)	0.03	0.24 (0.44)	0.21 (0.43)	0.00
Female	0.57 (0.50)	0.56 (0.50)	0.54	0.57 (0.49)	0.57 (0.49)	0.92
Full-time status	0.80 (0.40)	0.71 (0.46)	0.00	0.82 (0.38)	0.75 (0.43)	0.00
Dependent	0.80 (0.40)	0.70 (0.46)	0.00	0.82 (0.38)	0.74 (0.43)	0.00
Seeking a Bachelor’s degree	0.99 (0.12)	0.96 (0.19)	0.00	1.00 (0.06)	1.00 (0.04)	0.21
State Resident	0.74 (0.44)	0.84 (0.36)	0.00	0.76 (0.43)	0.86 (0.35)	0.00
Attends a private institution	0.54 (0.50)	0.35 (0.48)	0.00	0.55 (0.50)	0.34 (0.47)	0.00
Very selective institution	0.34 (0.47)	0.25 (0.43)	0.00	0.35 (0.50)	0.34 (0.47)	0.00
Moderately selective institution	0.53 (0.50)	0.58 (0.49)	0.00	0.55 (0.50)	0.62 (0.49)	0.00
Minimally selective institution	0.09 (0.29)	0.10 (0.30)	0.28	0.08 (0.27)	0.08 (0.27)	0.86
Open admissions institution	0.04 (0.18)	0.03 (0.25)	0.00	0.02 (0.15)	0.04 (0.19)	0.00
Observations	7,680	46,390	–	7,260	40,620	–

Notes: This table displays summary statistics. The left panel includes those of the “full” sample. Column 1 and 2 contains the means with standard deviations in parentheses for loan recipients and non-recipients, respectively. Column 3 contains the *p*-values for the difference in means. The right panel is arranged similarly and reflects the sample used for matching. Financial need is expressed in thousands and 2016 USD. Sample sizes are rounded to the nearest tens per the IES’s data security requests.

SOURCE: 2003-2004, 2007-2008, 2011-2012, & 2015-2016 NPSAS, and 2008-2009 & 2016-2017 B & B.

Table 3.2: Majors Among Perkins Loan Recipients

Major	Recipients	Non-recipients	<i>p</i> -value for diff.
Undeclared or not in a degree program	0.068	0.039	0.000
Humanities	0.126	0.121	0.147
Social/behavioral sciences	0.121	0.116	0.099
Life sciences	0.138	0.129	0.035
Physical sciences	0.026	0.024	0.024
Math	0.017	0.015	0.175
Computer/information science	0.026	0.036	0.000
Engineering	0.082	0.074	0.014
Education	0.081	0.089	0.027
Business/management	0.120	0.138	0.000
Health	0.082	0.096	0.000
Vocation/technical	0.010	0.010	0.166
Other technical/professional	0.103	0.113	0.007
<i>Total Count</i>	<i>7,680</i>	<i>46,390</i>	–

Notes: This table displays the shares of Perkins loan recipients and non-recipients across different college majors in 12 categories. The total counts in the bottom row are rounded to the nearest tens per the IES's data security requests. All shares are based on these rounded values.

SOURCE: 2008-2009 and 2016-2017 B & B.

Table 3.3: Choice of Occupation Among Perkins Loan Recipients

Occupation	Recipients	Non-Recipients	<i>p</i> -value for diff.
Primary or Secondary School Teacher	0.080	0.090	0.175
Military Service	0.030	0.040	0.928
Nursing	0.033	0.041	0.120
Medical Technician	0.027	0.020	0.163
Law Enforcement or Correctional Officer	0.003	0.004	0.825
Child or family services	0.007	0.004	0.528
Other – does not qualify for loan forgiveness	0.857	0.838	0.253
<i>Total Count</i>	<i>1,500</i>	<i>9,440</i>	–

Notes: This table displays the shares of Perkins loan recipients and non-recipients across different occupational categories that could qualify the former group of students for loan forgiveness. The total counts in the bottom row are rounded to the nearest tens per the IES's data security requests. All shares are based on these rounded values.

SOURCE: 2008-2009 and 2016-2017 B & B.

Table 3.4: Results – Effects of Perkins Loans

	OLS				Matching	
	$need_i > 0$ (1)	$need_i > 5,000$ (2)	$need_i > 10,000$ (3)	$need_i > 15,000$ (4)	Simple (5)	Bias- corrected (6)
Worked at all	-0.0515*** (0.00675)	-0.0524*** (0.00684)	-0.0523*** (0.00741)	-0.0480*** (0.00891)	-0.0569*** (0.00874)	-0.0561*** (0.00874)
Hours worked while enrolled	-1.697*** (0.185)	-1.697*** (0.188)	-1.659*** (0.200)	-1.347 (0.234)	-1.563*** (0.232)	-1.538*** (0.232)
Standardized GPA	-0.117*** (0.0134)	-0.118*** (0.0136)	-0.118*** (0.0149)	-0.115*** (0.0179)	-0.0116*** (0.0176)	-0.116*** (0.0176)
STEM Major	0.00774 (0.00716)	0.00582 (0.00730)	0.00600 (0.00793)	-0.000249 (0.00935)	0.00378 (0.00874)	0.00372 (0.00874)
Loan forgiveness occupation	-0.00165 (0.0108)	0.00106 (0.0111)	0.00153 (0.0125)	0.00378 (0.0153)	-0.000379 (0.0147)	-0.000632 (0.147)
Observations	54,080	47,470	34,350	21,820	14,520	14,520

Notes: This table displays the estimates of the average Perkins loan effects on all outcomes using OLS (Columns 1-4) and matching (Columns 5 and 6). Each row corresponds to a particular outcome variable. Also, each of Columns 1-4 corresponds to a sample restriction on financial need. The matching estimates incorporate all students with financial need above 0. The analyses on loan forgiveness occupation choice uses the B & B sub-sample. Thus, the corresponding sub-sample sizes are: 11,040, 9,600, 6,780, and 4,500 for Columns 1-4, respectively. Similarly, matching estimates of Columns 5 and 6 uses a trimmed B & B sub-sample of size 9,080. One star indicates a 10% significance level, two stars indicate a 5% level, and three stars indicate a 1% level. Standard errors in Columns 1-4 are clustered by institution-by-survey wave cells. Those of Columns 5 and 6 are Abadie-Imbens robust standard errors. Sample sizes are rounded to the nearest tens per the IES's data security requests.

SOURCE: 2003-2004, 2007-2008, 2011-2012, and 2015-2016 NPSAS, and 2008-2009 and 2016-2017 B & B.

Table 3.5: Summary Statistics for Covariates in the Matched Sample

	Recipients (1)	Non- recipients (2)	diff. <i>p</i> -val. (3)
Financial need	19.34 (10.49)	19.23 (10.27)	0.52
Underrepresented Minority	0.24 (0.44)	0.24 (0.44)	1.00
Female	0.57 (0.49)	0.57 (0.49)	1.00
Full-time status	0.82 (0.38)	0.82 (0.38)	1.00
Dependent	0.82 (0.38)	0.82 (0.38)	1.00
Seeking a Bachelor's degree	1.00 (0.06)	1.00 (0.06)	1.00
State Resident	0.76 (0.43)	0.76 (0.43)	1.00
Attends a private institution	0.55 (0.47)	0.55 (0.47)	1.00
Very selective institution	0.35 (0.50)	0.35 (0.50)	1.00
Moderately selective institution	0.55 (0.50)	0.55 (0.50)	1.00
Minimally selective institution	0.08 (0.27)	0.08 (0.27)	1.00
Open admissions institution	0.02 (0.15)	0.02 (0.15)	1.00
Observations	7,260	7,260	–

Notes: This table displays summary statistics for all covariates in the matched sample. Column 1 contains the means with standard deviations in parenthesis for Perkins loan recipients. Columns 2 contains those of non-recipients. Column 3 contains the *p*-values for the difference in means between these two groups. Financial need is expressed in thousands and 2016 USD. Sample sizes are rounded to the nearest tens per the IES's data security requests.

SOURCE: 2003-2004, 2007-2008, 2011-2012, and 2015-2016 NPSAS, and 2008-2009 and 2016-2017 B & B.

Appendix A

Additional Material for Chapter 1

A.1 Sample Details

Table A1.1: List of Test-Optional Liberal Arts Colleges

Year	School	State	Year	School	State
2002	Mount Holyoke College	MA	2014	Trinity College	CT
2004	Pitzer College	CA	2014	Beloit College	WI
2005	St. Lawrence University	WI	2014	Wesleyan University	CT
2005	Knox College	IL	2015	University of Puget Sound	WA
2005	Juniata College	PA	2015	The College of Idaho	ID
2005	Lawrence University	NY	2015	Kalamazoo College	OH
2005	College of the Holy Cross	MA	2015	Emanuel College	MA
2006	Susquehanna University	PA	2015	Transylvania University	PA
2006	Franklin & Marshall College	PA	2015	Allegheny College	PA
2006	Drew College	NJ	2016	Whittier College	CA
2006	Bennington College	VT	2016	Skidmore College	NY
2006	Gustavus Adolphus College	MN	2016	Warren Wilson College	NC
2007	Lake Forest College	IL	2016	Whitman College	WA
2007	Gettysburg College	PA	2016	Willamette University	OR
2007	Wittenberg University	OH	2016	Houghton University	NY
2007	Hobart William Smith College	NY	2016	Cornell College	IA
2007	Denison University	OH	2016	Presbyterian College	SC
2008	Stonehill College	MA	2017	Hanover College	IN
2008	Guilford College	NC	2017	Linfield University	OR
2008	Goucher College	MD	2017	Wells College	NY
2008	Augustana College	IL	2017	Ripon College	WI
2009	Albright College	PA	2017	Bloomfield College	NJ
2009	Agnes Scott College	GA	2017	Wofford College	SC
2009	Smith College	MA	2018	Doane University	NE
2009	Ursinus College	PA	2018	Birmingham-Southern College	AL
2010	The University of the South	TN	2018	Austin College	TX
2010	Washington & Jefferson College	PA	2019	Spring Hill College	AL
2010	Lycoming College	PA	2019	Bucknell University	PA
2010	Saint Michael's College	VT	2019	Southwestern University	TX
2011	Saint Anselm College	NH	2019	Randolph College	VA
2012	Earlham College	IN	2019	Hendrix College	AK
2013	Ohio Wesleyan University	OH	2019	Monmouth College	NJ
2013	Washington College	MD	2019	DePauw University	IN
2014	Bryn Mawr College	PA			

Notes: This table provides a list of included in the sample. The “Year” column correspond to each institution’s year of adopting their test-optional policy. For example, Pitzer College’s policy was effective in the Fall 2004 round of admissions.

Table A1.2: Summary Statistics (selective colleges)

	Test- optional (1)	Test- requiring (2)	p-value of diff. (3)
First-time URM students	26.54 (20.10)	20.20 (21.81)	0.27
Fraction of first-time students that are URM	0.08 (0.06)	0.06 (0.05)	0.26
4-year URM grad rate	0.45 (0.17)	0.34 (0.21)	0.04
6-year URM grad rate	0.57 (0.18)	0.42 (0.21)	0.01
Tuition & Fees	26,617 (4,492)	22,681 (5,300)	0.00
Full-time enrollment	1,267 (469)	1,088 (494)	0.18
Institutional grants per FTE	10.16 (5.29)	19.47 (35.11)	0.16
E & R expenditures per FTE	29.25 (22.80)	49.47 (73.85)	0.16
College prep courses not considered	0.04 (0.20)	0.00 (0.00)	0.32
College prep courses recommended	0.29 (0.46)	0.63 (0.49)	0.01
College prep courses required	0.67 (0.48)	0.37 (0.49)	0.03
Observations	24	30	–

Notes: This table is equivalent to Table 1.1, except it solely reflects colleges within the sample that are considered to be “selective” in admissions. The sample size varies slightly across each variable due to non-reporting.

Table A1.3: Summary Statistics (highly selective colleges)

	Test- optional (1)	Test- requiring (2)	p-value of diff. (3)
First-time URM students	33.49 (20.36)	35.44 (26.61)	0.69
Fraction of first-time students that are URM	0.08 (0.05)	0.09 (0.05)	0.58
4-year URM grad rate	0.57 (0.19)	0.62 (0.19)	0.24
6-year URM grad rate	0.66 (0.14)	0.71 (0.17)	0.14
Tuition & Fees	32,340 (5,028)	31,000 (7,287)	0.29
Full-time enrollment	1,613 (631)	1,536 (773)	0.60
Institutional grants per FTE	9.45 (4.43)	20.31 (47.32)	0.10
E & R expenditures per FTE	27.95 (14.95)	81.72 (201.43)	0.06
College prep courses not considered	0.02 (0.15)	0.04 (0.19)	0.67
College prep courses recommended	0.40 (0.49)	0.62 (0.49)	0.03
College prep courses required	0.56 (0.50)	0.33 (0.47)	0.02
Observations	43	52	–

Notes: This table is equivalent to Table 1.1, except it solely reflects colleges within the sample that are considered to be “highly selective” in admissions. The sample size varies slightly across each variable due to non-reporting.

A.2 Results without Control Variables

This section reproduces the estimates for Tables 1.3-1.5, but it excludes the use of control variables while retaining the institution and academic year fixed effects. The results for these exercises are provided in Tables A2.1-A2.3. In short, the point estimates are similar to the ones in Tables 1.3-1.5. Thus, the results are not sensitive to the exclusion of control variables (i.e., these variables merely improve the precision of the TWFE estimates).

Table A2.1: Primary Results – without control variables

	Logged Number of first-time URM students (1)	Fraction of first-time students that are URM (2)	4-year URM grad rate (3)	4-year URM grad rate (4)
Test-optional	0.123** (0.0406)	0.0191** (0.00644)	-0.00272 (0.0125)	0.00580 (0.0119)
Observations	2,827	2,831	2,790	2,819

Notes: This table displays the results from estimating equation (1.1) on all outcome variables, but with a specification excluding control variables. The row labeled with “Test-optional” contains the point estimate of TWFE coefficient. Each column corresponds to an outcome variable. Columns (3) and (4) correspond to graduation outcomes, so the regressions for those use a lagged treatment indicator (i.e., $P_{i,t-6}$). One star indicates a 5% significance level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

Table A2.2: Effects of the Policy across Selectivity – without control variables

	Logged Number of first-time URM students (1)	Fraction of first-time students that are URM (2)	4-year URM grad rate (3)	6-year URM grad rate (4)
Test-optional	0.154* (0.0690)	0.0247* (0.0121)	-0.0678** (0.0153)	-0.0476** (0.0149)
Highly Selective × Test-optional	-0.0422 (0.0831)	-0.00792 (0.0139)	0.0820** (0.0211)	0.0662** (0.0200)
Observations	2,827	2,831	2,790	2,819

Notes: This table displays the point estimate from equation (1.2) without control variables. The row labeled with “Test-optional” contains the point estimate for the coefficient of the treatment indicator of having the policy in place. The row labeled “Highly selective × Test-optional” contains the point estimate for the coefficient of the interaction term between the treatment indicator and another indicator for being highly selective in admissions. Each column corresponds to an outcome variable. The point estimates for the coefficients of the control variables are not reported in this table. One stars indicate a 5% level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

Table A2.3: Effects of the Policy by Adoption Timing – without control variables

	Logged Number of first-time URM students (1)	Fraction of first-time students that are URM (2)
Early adopter	0.121* (0.0586)	0.0215* (0.0106)
Late adopter	0.127* (0.0510)	0.0179* (0.00791)
<i>p</i> -value	.939	.799
Observations	2,827	2,831

Notes: This table displays the point estimate from equation (1.3), but with a specification that excludes control variables. The row labeled “Early adopter” corresponds to the estimated effect of the policy on institutions that dropped the test requirement early. Similarly, the row labeled “Later adopter” corresponds to the effect on institutions adopting the policy later. Each column corresponds to a first-time URM outcome variable. The *p*-values on the bottom row reflect the test of $\beta_1 = \beta_2$. One star indicates a 5% level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

A.3 Propensity-Trimmed Sample

This section replicates the estimates of equations (1.1) and (1.2) using a sample derived from a propensity trimming technique suggested by Crump et al. (2009).¹ A propensity-trimmed sample is constructed in a few steps. First, this study estimates a logit model of the probability that an institution adopted the policy as a function of the 2001-2002 (pre-treatment) observations of the covariates used in equation (1.1), as well as the fraction of students that are from a URM background (i.e., regardless of freshman status), and the 6-year URM graduation rate. Motivated by Section 1.5.2, this logit model is estimated for selective and highly selective institutions. Second, it calculated the fitted values of these regressions, which serve as each institution's estimated propensity for adopting the test-optional policy. Third, it discards all institutions whose propensity score lies outside of $[0.1, 0.9]$. Therefore, this method effectively excludes institutions that have virtually no probability of adopting as well as those that will almost certainly do so.

Figure A3.1 contains the densities of the institutions' propensity scores, which were calculated by the logit regressions, across treatment groups and selectivity levels. Furthermore, the results of these regressions can be found in Table A3.1.² Interestingly, they show that the coefficients for overall campus diversity and the 6-year URM graduation rate are insignificant. Therefore, these variables may not be strong predictors of adopting the test-optional policy, regardless of selectivity.

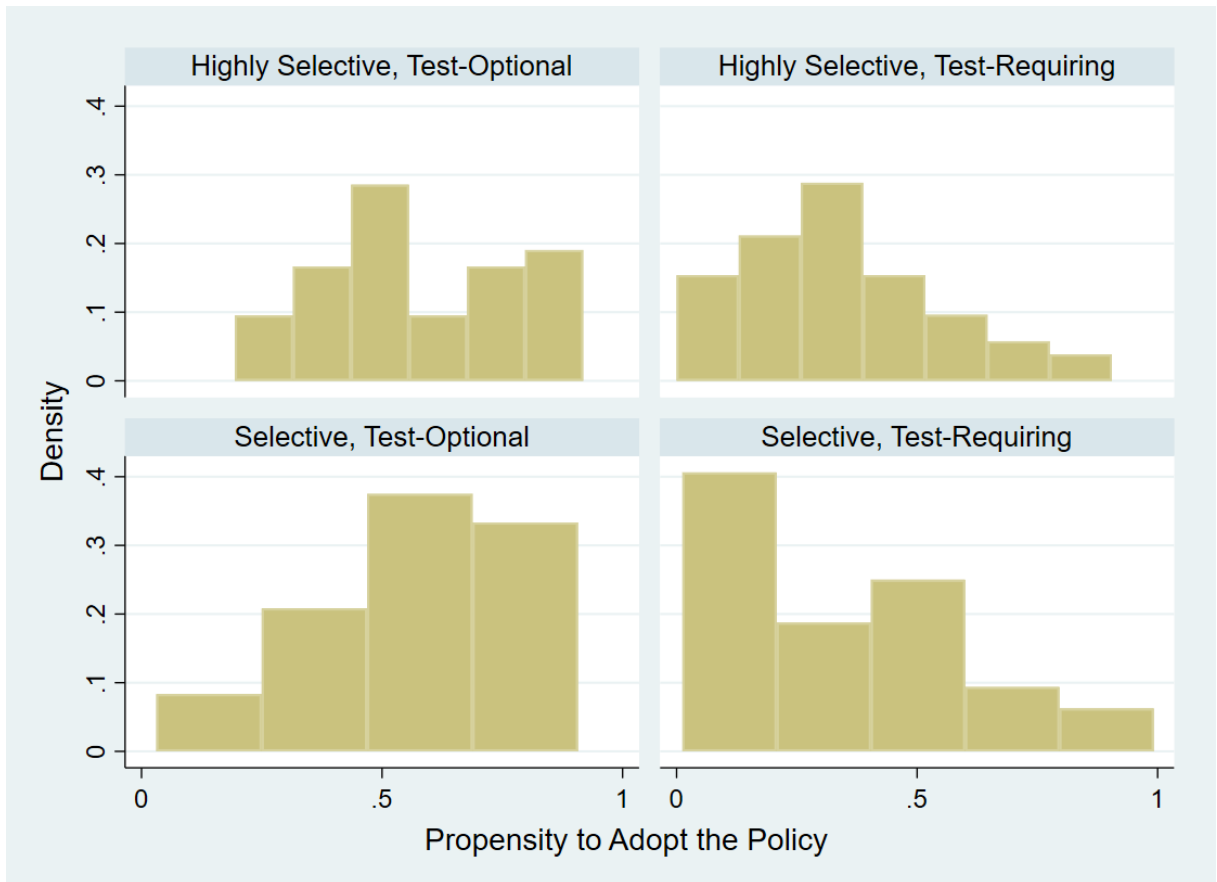
The resulting propensity-trimmed sample contains 130 institutions, of which 44 and 86 are selective and highly selective, respectively (i.e., 19 institutions are dropped due to

¹This technique was originally proposed to address a lack of overlap in the distribution of outcome and control variables. This issue could lead to estimates being substantially biased and having large variances.

²None of the selective institutions required college prep courses in 2001, the logit regression does not estimate a coefficient for the corresponding indicator variable.

having propensities less than 0.10 or greater than 0.90). In each of these sub-samples, at least 48% of institutions are test-optional. The summary statistics for the full sample, selective institutions, and highly selective institutions can be found in Tables A3.2, A3.4, and A3.5. Compared to Table 1.4, the differences in mean for the 4-year and 6-year URM graduation rates among selective colleges are relatively balanced. Although the differences are sizable (e.g., approximately 0.08 for the 6-year URM graduation rate), they are statistically indistinguishable from zero. Finally, Tables A3.3 and A3.6 replicate Tables A1.2 and A1.3 the propensity-trimmed sample. The point estimates between these sets of tables are comparable in magnitude. Thus, the estimates from Tables 1.3 and 1.4 are not biased by the inclusion of institutions that were either likely to adopt the test-optional policy or were unlikely to do so.

Figure A3.1



Notes: This figure displays the distribution of institutions' propensity to adopt the test-optional policy across selectivity and treatment group (i.e., test-optional versus test-requiring schools).

Table A3.1: Logit Regressions

	Selective (1)	Highly selective (2)
Fraction of students that are URM	7.629 (5.042)	6.097 (7.195)
6-year graduation rate	3.329 (2.099)	-1.565 (2.003)
Logged tuition & fees	7.962 (4.943)	1.455 (1.886)
Logged full time enrollment	-2.347 (1.816)	1.042 (0.597)
Logged institutional grants per FTE	-1.352 (1.221)	2.768* (1.202)
Logged E & R expenditures per FTE	-1.993 (1.927)	-2.329* (1.028)
College prep courses recommended	-0.211 (0.775)	-0.940 (1.629)
College prep courses required		0.136 (1.635)
Observations	54	95

Notes: This table contains the results of the logistic regressions used to construct the propensity trimmed sample. These regressions incorporate 2001-2002 (pre-treatment) observations. They were run separately for selective and highly selective colleges. The point estimates for the former and latter are in columns (1) and (2), respectively. One star corresponds to a 5% significance level while two stars correspond to a 1% level. Standard errors are in parenthesis and clustered by institution.

Table A3.2: Summary Statistics

	Test- optional (1)	Test- requiring (2)	p-value of diff. (3)
First-time URM students	31.19 (20.77)	30.53 (23.23)	0.87
Fraction of first-time students that are URM	0.08 (0.05)	0.08 (0.05)	0.68
4-year URM grad rate	0.53 (0.19)	0.55 (0.20)	0.64
6-year URM grad rate	0.63 (0.16)	0.63 (0.20)	0.98
Tuition & Fees	30,351.33 (5,396.33)	29,355.47 (6,036.80)	0.32
Full-time enrollment	1,494.63 (606.94)	1,475.33 (677.58)	0.86
Institutional grants per FTE	9.71 (4.80)	18.53 (43.90)	0.11
E & R expenditures per FTE	28.49 (18.34)	62.10 (167.21)	0.11
College prep not considered	0.03 (0.18)	0.03 (0.17)	0.98
College prep recommended	0.36 (0.48)	0.58 (0.50)	0.01
College prep required	0.59 (0.50)	0.38 (0.49)	0.01
Observations	64	66	–

Notes: Columns (1) and (2) contain summary statistics for test-optional and test-requiring institutions using the 2001-2002 (i.e., pre-treatment) observations. However, this table incorporates the propensity-trimmed sample. Standard deviations are in parenthesis. Column (3) contains the p -values from the difference means between these two groups. These p -values are clustered by institution. None of the variables are logged. Therefore, the means for tuition & fees, E & R expenditures per FTE, and institutional grants per FTE are in terms of dollars. The sample size varies slightly across each variable due to non-reporting.

Table A3.3: Primary Results under Propensity-Trimmed Sample

	Logged Number of first-time URM students (1)	Fraction of first-time students that are URM (2)	4-year URM grad rate (3)	4-year URM grad rate (4)
Test-optional	0.124** (0.0413)	0.0169** (0.00634)	0.00773 (0.0145)	0.00843 (0.0135)
Observations	2,459	2,463	2,428	2,454

Notes: This table displays the results from estimating equation (1.1) on all outcome variables. However, these estimates are based on the propensity-trimmed sample. The row labeled with “Test-optional” contains the point estimate of the TWFE coefficient. Each column corresponds to an outcome variable. Columns (3) and (4) correspond to graduation outcomes, so the regressions for those use a lagged treatment indicator (i.e., $P_{i,t-6}$). One star indicates a 5% level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

Table A3.4: Summary Statistics (selective colleges)

	Test- optional (1)	Test- requiring (2)	p-value of diff. (3)
First-time URM students	26.95 (20.90)	24.09 (24.38)	0.68
Fraction of first-time students that are URM	0.08 (0.06)	0.07 (0.06)	0.52
4-year URM grad rate	0.45 (0.18)	0.42 (0.17)	0.59
6-year URM grad rate	0.58 (0.18)	0.50 (0.17)	0.16
Tuition & Fees	26,584.54 (3,764.73)	24,644.84 (4,025.50)	0.11
Full-time enrollment	1,279.45 (488.19)	1,219.14 (448.53)	0.67
Institutional grants per FTE	10.19 (5.42)	16.80 (33.01)	0.36
E & R expenditures per FTE	29.00 (23.83)	40.03 (67.37)	0.47
College prep not considered	0.05 (0.21)	0.00 (0.00)	0.32
College prep recommended	0.27 (0.46)	0.55 (0.51)	0.07
College prep required	0.68 (0.48)	0.45 (0.51)	0.13
Observations	22	22	–

Notes: This table is equivalent to Table A1.2, but it reflects the “selective” colleges from the propensity-trimmed sample. The sample size varies slightly across each variable due to non-reporting.

Table A3.5: Summary Statistics (highly selective colleges)

	Test- optional (1)	Test- requiring (2)	p-value of diff. (3)
First-time URM students	33.40 (20.60)	33.75 (22.21)	0.94
Fraction of first-time students that are URM	0.08 (0.05)	0.08 (0.05)	0.98
4-year URM grad rate	0.58 (0.19)	0.61 (0.18)	0.36
6-year URM grad rate	0.66 (0.14)	0.70 (0.17)	0.26
Tuition & Fees	32,324.42 (5,087.73)	31,710.78 (5,490.42)	0.59
Full-time enrollment	1,607.33 (637.27)	1,603.43 (738.50)	0.98
Institutional grants per FTE	9.45 (4.49)	19.40 (48.77)	0.18
E & R expenditures per FTE	28.22 (15.02)	73.13 (199.18)	0.14
College prep not considered	0.02 (0.15)	0.05 (0.21)	0.59
College prep recommended	0.40 (0.50)	0.59 (0.50)	0.09
College prep required	0.55 (0.50)	0.34 (0.48)	0.05
Observations	42	44	–

Notes: This table is equivalent to Table A1.3, but it reflects the highly “selective colleges” from the propensity-trimmed sample. The sample size varies slightly across each variable due to non-reporting.

Table A3.6: Effects of the Policy across Selectivity (under Propensity-Trimmed Sample)

	Logged Number of first-time URM students (1)	Fraction of first-time students that are URM (2)	4-year URM grad rate (3)	6-year URM grad rate (4)
Test-optional	0.160* (0.0723)	0.0243* (0.0117)	-0.0543* (0.0229)	-0.0543* (0.0216)
Highly selective \times Test-optional	-0.0230 (0.0857)	-0.00578 (0.0138)	0.0707* (0.0272)	0.0730** (0.0253)
Observations	2,573	2,577	2,542	2,568

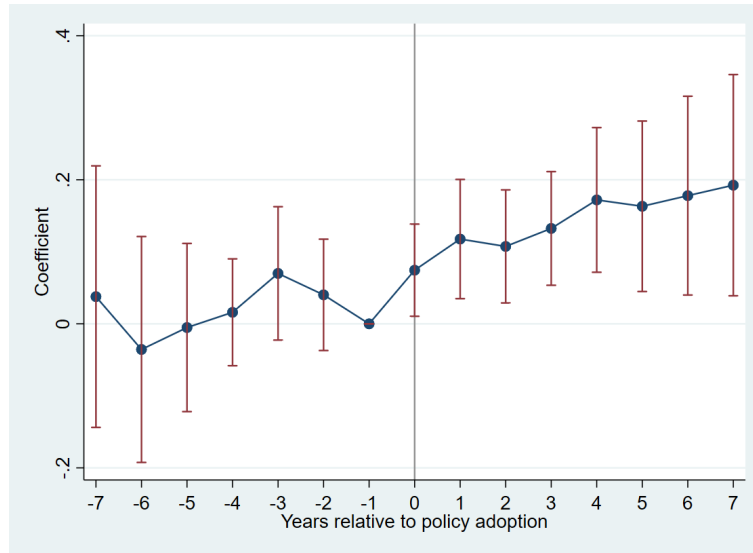
Notes: This table displays the point estimate from equation (1.2), which distinguishes institutions by their degree of admission selectivity. However, these correspond to the propensity-trimmed sample. The row labeled with “Test-optional” contains the point estimate for the coefficient of the treatment indicator of having the policy in place. The row labeled “Highly selective \times Test-optional” contains the point estimate for the coefficient of the interaction term between the treatment indicator and another indicator for being highly selective in admissions. Each column corresponds to an outcome variable. The point estimates for the coefficients of the control variables are not reported in this table. One star indicates a 5% significance level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

A.4 Heterogeneity Robust Estimation

This section replicates Figures 1.2-1.5 using alternative estimators suggested by de Chaisemartin and D’Haultfoeuille (2021). These estimators, denoted as DID_ℓ , are robust to treatment heterogeneity and dynamic effects. So, in place of leads and lags, this study estimates dynamic and placebo effects, respectively. Therefore, a joint placebo test serves as an analogous and robust test for the common trends assumption. Figures A4.1-A4.4 contain the plotted placebo and dynamic effect estimates across “event time” (i.e., the period relative to adoption) for all outcome variables. The p-values for the joint placebo tests are displayed in Table D1 across each outcome variable. Similar to the joint F-test of the leads, these tests also fail to find strong evidence that at least one placebo is statistically different from 0. Thus, the common trends assumption is presumed to hold under this specification.

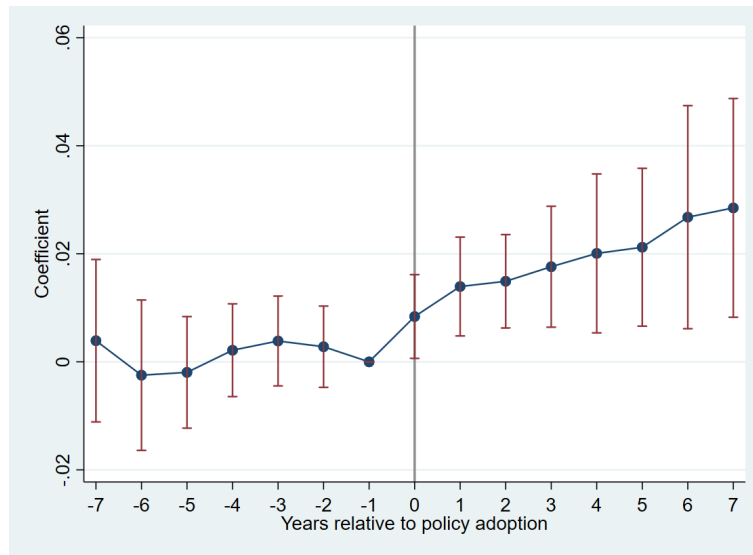
This section also re-estimate the URM enrollment effects from columns (1)-(4) of Table 1.3 using the average of the dynamic DID_ℓ estimates discussed above. These serve as analogues to the original two-way fixed effects estimates from Table 1.3. The average effect for each outcome is reported in Table A4.2, with bootstrapped standard errors displayed in parenthesis. The point estimates are mostly comparable to the ones from the primary two-way fixed effects specification, so, therefore, heterogeneous treatment effects may not be a significant issue. Although the DID_ℓ point estimates for the URM graduation rates differ from that of TWFE, they are imprecise.

Figure A4.1



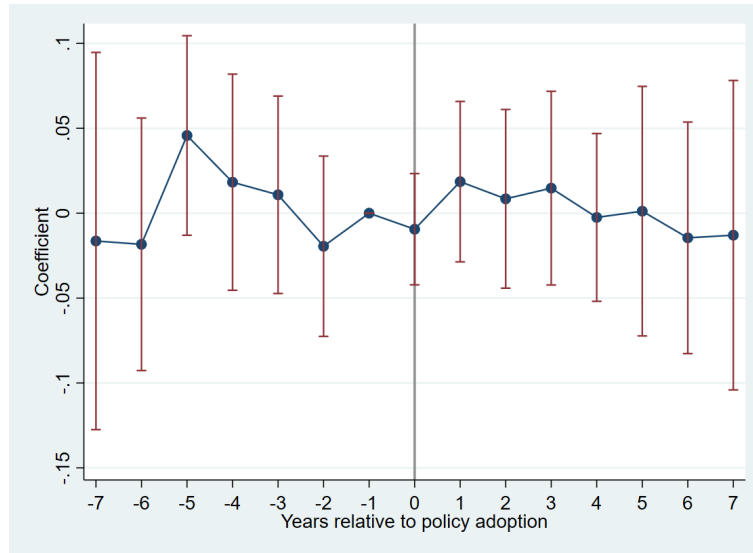
Notes: This figure illustrate the computed placebo and dynamic effects of the test-optional policy on the logged number of URM students enrolling for the first-time. The placebo and dynamic effects are analogous to the leads and lags of an event-study. These estimates are plotted across “event time.” They are represented by the dots. The accompanying bands represent the 95% confidence intervals of these estimates, which were bootstrapped across 100 replications. All proceeding figures are arranged similarly, but they reflect different outcome variables.

Figure A4.2



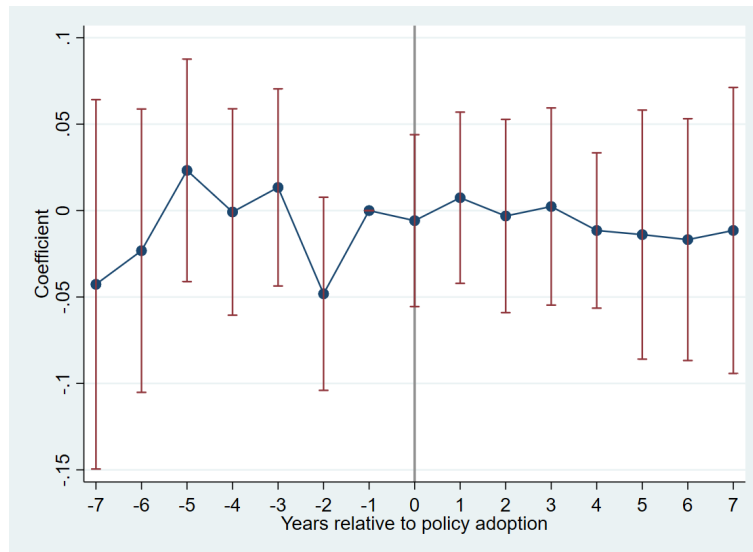
Notes: This figure illustrates the computed placebo and dynamic effects of the test-optional policy on the fraction of first-time students enrolling at liberal arts colleges that are of URM status.

Figure A4.3



Notes: This figure illustrates the computed placebo and dynamic effects of the test-optional policy on the 4-year graduation rate for URM students.

Figure A4.4



Notes: This figure illustrates the computed placebo and dynamic effects of the test-optional policy on the 6-year graduation rate for URM students.

Table A4.1: Results of Joint Placebo Test

	<i>p</i> -value for joint placebo test
a) Logged number of first-time URM Students	0.561
b) Fraction of first-time students that are URM	0.913
c) 4-year URM graduation rate	0.212
d) 6-year URM graduation rate	0.192

Notes: This table displays the *p*-values for the joint placebo tests across all outcome variables of interest. The point estimates of the placebo effects are displayed in Figures A4.1-A4.4.

Table A4.2: Heterogeneity-Robust Estimates

	DID_ℓ estimate (1)	TWFE estimate (2)
a) Logged number of first-time URM students	0.133 (0.0358)	0.125 (0.0419)
b) Fraction of first-time students that are URM	0.0174 (0.00453)	0.0188 (0.00628)
c) 4-year URM grad rate	-2.32e-05 (0.0244)	0.00247 (0.0142)
d) 6-year URM grad rate URM	-0.00507 (0.0255)	0.00450 (0.0131)

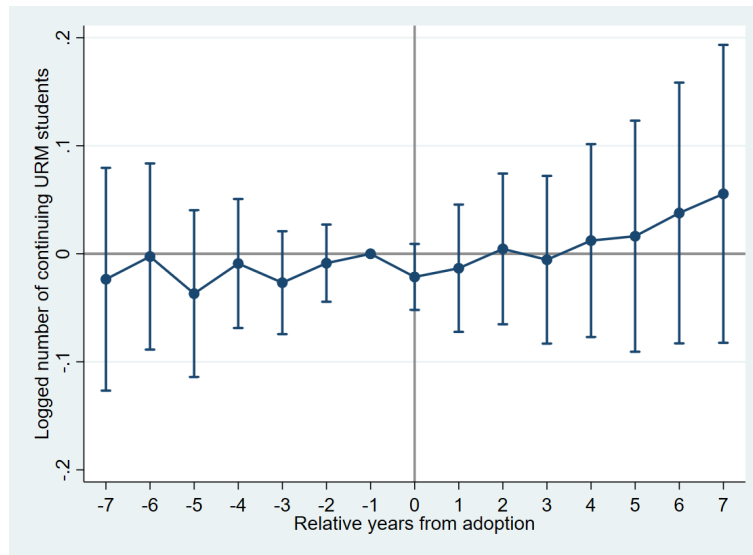
Notes: This table reflects the DID_ℓ estimates for each outcome variable. Thus, these estimates are robust to heterogeneous and dynamic treatment effects. Column (1) contain the point estimates with bootstrapped standard errors in parenthesis. These standard errors are iterated across 100 replications. Column (2) displays the two-way fixed effects estimates from Table 1.3 for comparison. Clustered standard errors are in parenthesis. Sample sizes across each regression are roughly 2,820, but they somewhat vary due to some non-reporting for some left-handed variables.

A.5 Effects of the Policy on Non-first-time students

The effects of the policy should have little impact on the enrollment outcomes of URM students that are not in their first year of college (i.e., non-freshman). This study estimates equation (1.1) using the outcome variables for non-freshman URM enrollment to investigate this possibility, i.e., the logged number of non-freshman URM students and the fraction of non-freshman students that are of a URM background. Figures A5.1 and A5.2 illustrate dynamic effects the policy on these outcomes, and they do indicate any strong evidence of pre-treatment trends. Although the dynamic effects increase across event time in Figure A5.2, they are not significant. Columns (1) and (2) of Table A5.2 displays the point estimates of estimating equation (1.1). Column (1) indicates that the policy has a negligible effect on the volume of non-freshman URM students. Interestingly, however, this policy has a small and significant impact on the fraction of non-freshmen of a URM background.

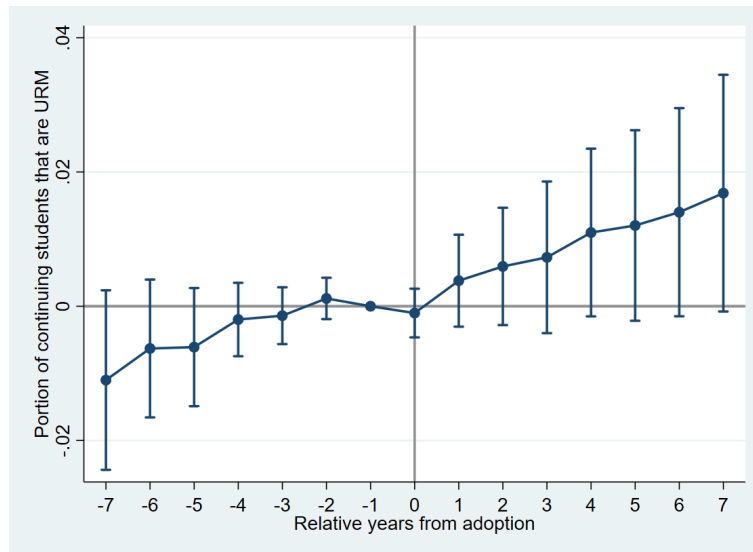
To investigate this phenomenon further, this section re-estimates equation (1.2) using these outcome variables to distinguish the effects between selective and highly selective colleges. The respective estimates are displayed in columns (1.3) and (1.4). It shows that the policy has little impact on the volume of non-freshman URM students at selective and highly selective colleges. Similarly, the policy has little impact on the fraction of non-freshman students from a URM background for both types of colleges. In short, there is insufficient evidence that the policy increased this fraction at each type of college. But it is possible that pooled effect on this outcome merely reflects sufficient statistical power.

Figure A5.1



Notes: This figure illustrate the dynamic effects on the logged number of non-first-time URM students. The p -value for the joint test of the leads is 0.450.

Figure A5.2



Notes: This figure illustrate the dynamic effects for the share of non-first-time students from a URM background. The p -value for the joint test of the leads is 0.269.

Table A5.1: Effects of the policy on non-first-time URM enrollment

	Logged Number of non-freshmen URM students (1)	Fraction of non-freshmen that are URM (2)	Logged Number of non-freshmen URM students (3)	Fraction of non-freshman that are URM (4)
Test-optional	0.0204 (0.0427)	0.0106* (0.00486)	-0.0155 (0.0705)	0.00985 (0.00895)
Highly-selective \times Test-optional			0.0626 (0.0866)	2.66e-05 (0.0107)
Observations	2,824	2,819	2,824	2,819

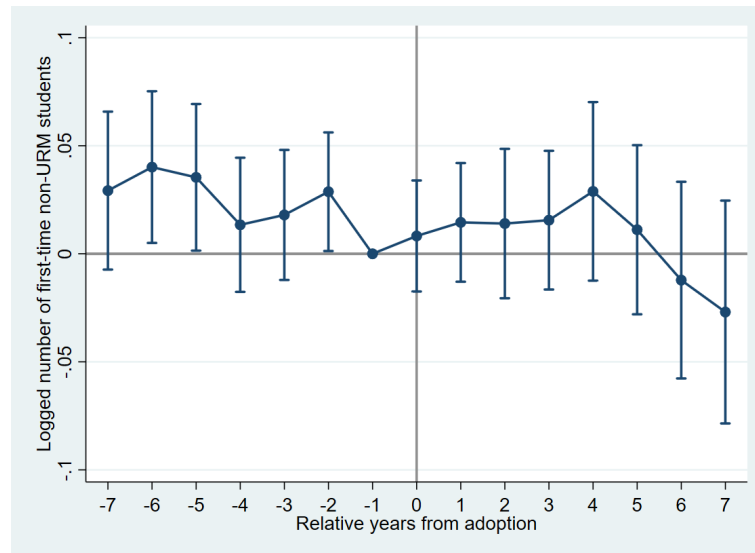
Notes: Columns (1) and (2) displays the point estimates the policy's effects from equation (1.1) while columns (3) and (4) display those of equation (1.2). Each column corresponds to an outcome variable. The row labeled with "Test-optional" contains the point estimate for the coefficient of the treatment indicator of having the policy in place. The row labeled "Highly selective \times Test-optional" contains the point estimate for the coefficient of the interaction term between the treatment indicator and another indicator for being highly selective in admissions. One star corresponds to a 5% significance level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

A.6 Enrollment of First-time Non-URM Students

This section discusses the effects of the test-optional policy on the logged number of first-time non-minority students. The purpose of this analysis is to show whether institutions within the sample are likely to implement the test-optional policy to merely boost their enrollment of students (i.e., regardless of racial background) rather than to improve their racial diversity. In this case, the policy should increase the overall number of non-URM first-time students concurrently with that of UMR students enrolling at these institutions.

Figure A6.1 displays the dynamic effects of test-optional admissions on the logged number of non-URM first-time students. It indicates that the parallel trends assumption is likely to be plausible ($p = .160$), although the point estimates for a few leads are significant. Column (1) of Table A6.1 displays the point estimate of the test-optional policy using equation (1.1). It suggests that within the sample as a whole, the policy has a small, negative impact on the volume of these students enrolling at these colleges (i.e., by 1.35%). However, this point estimate is insignificant. Column (2) displays the point estimates from estimating equation (1.2) to distinguish the enrollment effects across institution selectivity, and it shows that the enrollment of first-year non-URM students decreased by 22.4% as a result of the policy. This point estimate is also insignificant. Similarly, the policy has little impact on this outcome selective colleges ($p = .889$). All in all, these results suggest that most of the colleges in the sample are likely to be aiming to improve campus racial diversity rather than simply bolstering their enrollment levels.

Figure A6.1



Notes: This figure illustrates the dynamic effects on the logged number of non-URM first-time students. Coefficient estimates from equation (1.4) are plotted across event time. They are represented by the blue dots. The accompanying bands represent the 95% confidence intervals of these coefficients. The p -value for the joint test of the leads is 0.160.

Table A6.1: Effect of Test-Optional Policy on Logged First-time Enrollment

	Main Specification (1)	Selectivity Heterogeneity (2)
Test-optional	-0.0135 (0.0112)	-0.00618 (0.0200)
Highly Selective \times Test-optional		-0.0162 (0.0239)
Observations	2,820	2,820

Notes: This table displays the results from estimating equation (1.1) and (1.2) on the logged number of first-time students. The point estimates from each equation correspond to columns 1 and 2, respectively. The row labeled with “Test-optional” contains the point estimate for the coefficient of the treatment indicator of having the policy in place. The row labeled “Highly selective \times Test-optional” contains the point estimate for the coefficient of the interaction term between the treatment indicator and another indicator for being highly selective in admissions. Standard errors are in parenthesis and clustered by institution.

A.7 Inclusion of HBCU Institutions

The initial sample of institutions included three Historically Black Colleges & Universities (HBCUs): Morehouse College, Spelman College, and Fisk University. All of these are considered to be “selective” by the USNWR. However, they are excluded from the primary analyses. Tables A7.1 and A7.2 replicate the results from Tables 1.3 and 1.4 while including HBCUs. In Table A7.1, the point estimates for the effects of policy on URM enrollment outcomes (e.g., the fraction of freshman students that are URM) are slightly larger than the point estimates from Table 1.3. Similarly, in Table A7.2, the point estimates for these outcomes among selective colleges are also larger than the point estimates from Table 1.4. The larger point estimates could be attributed to the trends in these outcomes that these HBCU institutions follow across time within the panel. However, the difference in point estimates between selective and highly selective colleges remains statistically insignificant.

Table A7.1: Primary Results (with HBCUs)

	Logged Number of first-time URM students (1)	Fraction of first-time students that are URM (2)	4-year URM grad rate (3)	6-year URM grad rate (4)
Test-optional	0.140** (0.0428)	0.0218** (0.00667)	0.00123 (0.0140)	0.00609 (0.0129)
Observations	2,877	2,881	2,837	2,869

Notes: This table displays the results from estimating equation (1.1) on all outcome variables. This table, however, reflect the point estimates from a sample that includes HBCUs. The row labeled with “Test-optional” contains the point estimate of the TWFE coefficient. Each column correspond to an outcome variable. Columns 4-7 correspond to graduation outcomes, so the regressions for those use a lagged treatment indicator (i.e., $P_{i,t-6}$). One stars indicates a 5% significance level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression may vary slightly due to some non-reporting for some left-handed variables. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

Table A7.2: Effects of the Policy across Selectivity (with HBCUs)

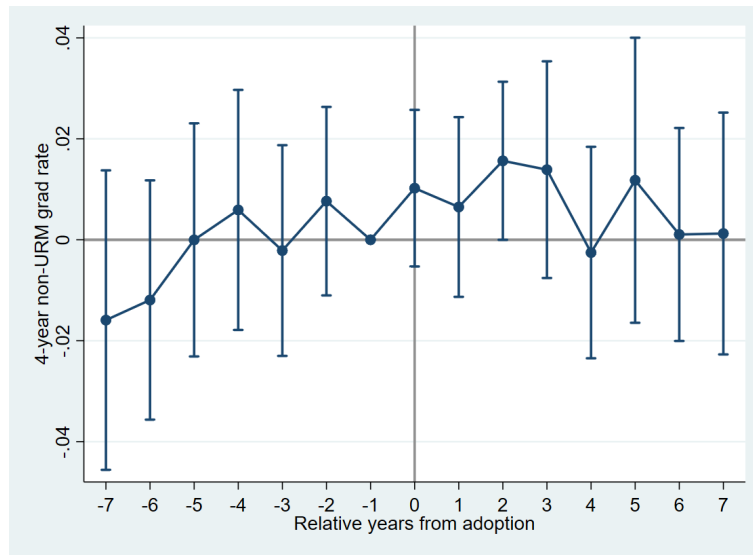
	Logged Number of first-time URM students (1)	Fraction of first-time students that are URM (2)	4-year URM grad rate (3)	6-year URM grad rate (4)
Test-optional	0.199** (0.0727)	0.0335* (0.0133)	-0.0508* (0.0204)	-0.0461* (0.0196)
Highly selective × Test-optional	-0.0785 (0.0872)	-0.0155 (0.0151)	0.0656* (0.0252)	0.0632** (0.0236)
Observations	2,877	2,881	2,837	2,869

Notes: This table displays the point estimate from equation (1.2), which distinguishes institutions by their degree of admission selectivity. However, these reflect a sample that includes HBCUs. The row labeled with “Test-optional” contains the point estimate for the coefficient of the treatment indicator of having the policy in place. The row labeled “Highly selective × Test-optional” contains the point estimate for the coefficient of the interaction term between the treatment indicator and another indicator for being highly selective in admissions. Each column corresponds to an outcome variable. The point estimates for the coefficients of the control variables are not reported in this table. One star indicates a 5% significance level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression slightly vary due to non-reporting for some left-handed variables.

A.8 Non-URM Graduation Rates

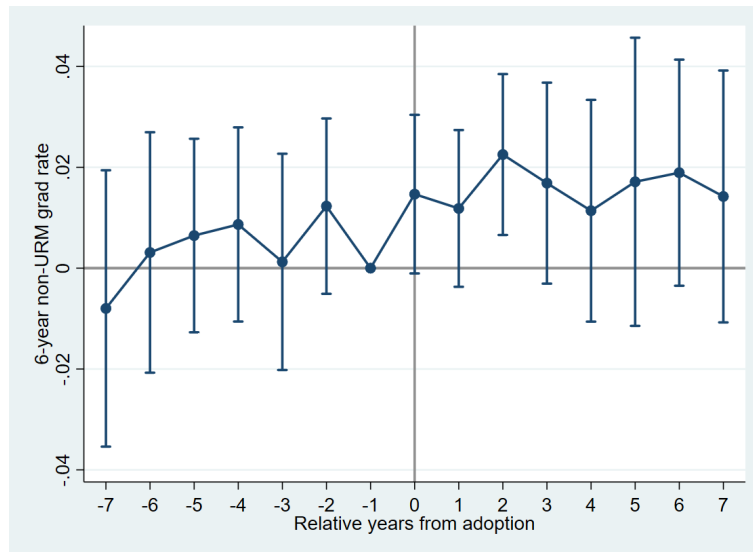
In this section, this study investigates whether the policy affects the graduation rates for non-URM students (e.g., White and Asian students). Figures A8.1 and A8.2 illustrate the dynamic effects of the policy on these graduation rates and shows little evidence for pre-treatment trends. Columns (1) and (2) of Table H1 shows that the policy has little impact on 4 and 6-year graduation rates for non-URM students throughout the sample as a whole. Furthermore, columns (3) and (4) indicate that neither selective nor highly selective colleges experience non-URM graduation rate effects as a result of this policy. This shows that the decline in the URM graduation rates at selective institutions are not driven by a drop in the overall campus graduation rates.

Figure A8.1



Notes: This figure illustrate the dynamic effects on the logged number of the 4-year graduation rate for non-URM students. The p -value for the joint test of the leads is 0.519.

Figure A8.2



Notes: This figure illustrate the dynamic effects for the share of the 6-year graduation rate for non-URM students. The p -value for the joint test of the leads is 0.239.

Table A8.1: Graduation rate of non-URM students

	4-year grad rate (1)	6-year grad rate (2)	4-year grad rate (3)	6-year grad rate (4)
Test-optional	0.000254 (0.00793)	0.00689 (0.00701)	-0.0187 (0.0192)	-0.00582 (0.0148)
Highly selective \times Test-optional			0.0253 (0.0210)	0.0148 (0.0169)
Observations	2,792	2,821	2,792	2,821

Notes: Columns (1) and (2) of this table results from estimating equation (1.1). Similarly, columns (3) and (4) results from estimating equation (1.2). Each column corresponds to an outcome variable. One stars indicates a 5% significance level and two stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution. Sample sizes across each regression may vary slightly due to some non-reporting for some left-handed variables.

A.9 Analysis at the State Level

Colleges included within the sample are located across 37 states. Of these, 29 contain at least one test-optional institution. However, some states, such as Pennsylvania, have multiple test-optional colleges. As discussed in Section 1.5.1, colleges may drop their test requirement when neighboring colleges within a state do so. For example, as shown in Table A1.1, Southwestern University in Texas dropped the test requirement one year after Austin College. Furthermore, several colleges from the same state switched to test-optional admissions simultaneously (e.g., Trinity College and Wesleyan University in 2015).

This study collapses the panel to the state level to investigate this possible phenomenon. It estimates a version of equation (1.1) displayed below:

$$y_{st} = \beta A_{st} + \mathbf{X}'_{st}\gamma + \theta_s + \lambda_t + \eta_{st}. \quad (\text{A.1})$$

Here, y_{st} either corresponds to the logged number of URM students in each state s or the share of students in each state from a URM background (i.e., these are state-level freshman URM enrollment outcomes). A_{st} corresponds to the share of students in each state that are enrolled at a test-optional college. Finally, θ_s and η_{st} are the state fixed effects and the stochastic error term, respectively. The covariates within \mathbf{X}_{st} are primarily state-level averages (e.g., logged average tuition & fees among colleges in state s).

The results of estimating equation (1.5) are displayed in Table A9.1. The point estimates for each state-level freshman URM enrollment outcome are, in short, noisy. Thus, there is little evidence that colleges self-select themselves for treatment to follow the steps of other colleges in the same state.

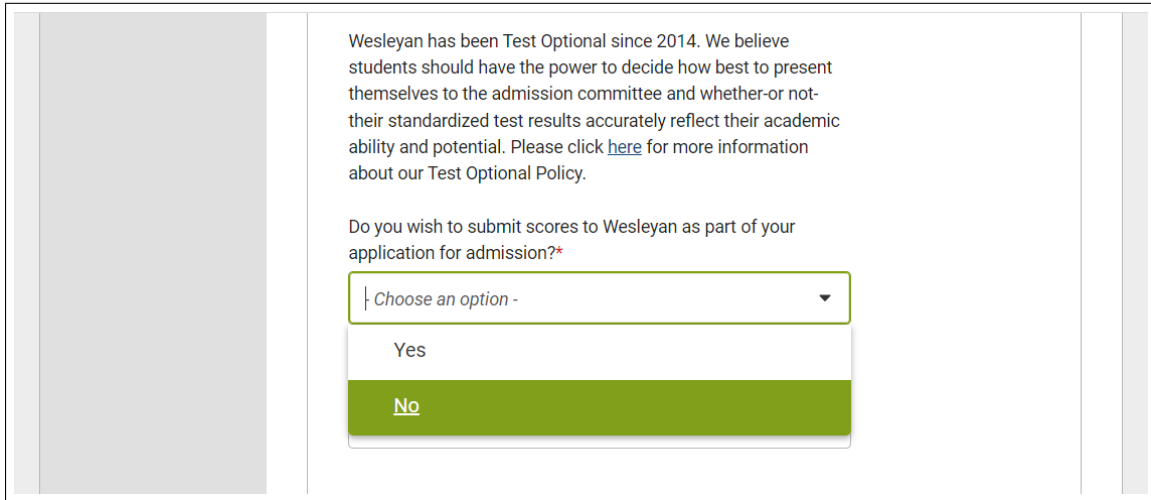
Table A9.1: State-level Effects of Test-Optional Admissions

	Logged number of URM (1)	Portion of first-time that are URM (2)
Share Test-Optional	0.115 (0.110)	0.00921 (0.0148)
Observations	703	703

Notes: This table results from estimating equation (1.5). Each column corresponds to an outcome variable related to freshman URM enrollment. Standard errors are in parenthesis and clustered by state.

A.10 Miscellaneous Figures and Tables

Figure A10.1



Wesleyan has been Test Optional since 2014. We believe students should have the power to decide how best to present themselves to the admission committee and whether-or-not their standardized test results accurately reflect their academic ability and potential. Please click [here](#) for more information about our Test Optional Policy.

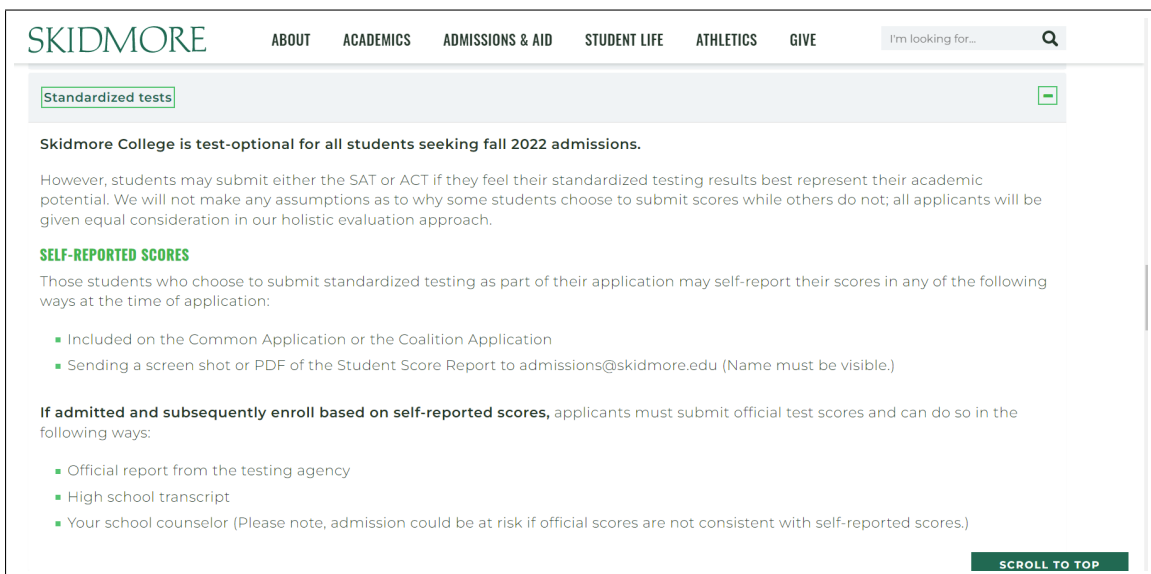
Do you wish to submit scores to Wesleyan as part of your application for admission?*

Choose an option -

- Yes
- No**

Notes: This figure is a screenshot of the Common Application page for Wesleyan University, a test-optional college. Similar to other test-optional institutions, the application for this college asks the applicant to choose whether they wish to submit their SAT or ACT test scores for admission consideration.

Figure A10.2



SKIDMORE ABOUT ACADEMICS ADMISSIONS & AID STUDENT LIFE ATHLETICS GIVE I'm looking for... Q

Standardized tests

Skidmore College is test-optional for all students seeking fall 2022 admissions.

However, students may submit either the SAT or ACT if they feel their standardized testing results best represent their academic potential. We will not make any assumptions as to why some students choose to submit scores while others do not; all applicants will be given equal consideration in our holistic evaluation approach.

SELF-REPORTED SCORES

Those students who choose to submit standardized testing as part of their application may self-report their scores in any of the following ways at the time of application:

- Included on the Common Application or the Coalition Application
- Sending a screen shot or PDF of the Student Score Report to admissions@skidmore.edu (Name must be visible.)

If admitted and subsequently enroll based on self-reported scores, applicants must submit official test scores and can do so in the following ways:

- Official report from the testing agency
- High school transcript
- Your school counselor (Please note, admission could be at risk if official scores are not consistent with self-reported scores.)

SCROLL TO TOP

Notes: This figure is a screenshot of the admissions page for a test-optional school, Skidmore College. They suggest that they will provide equitable consideration between applicants that choose to submit their SAT or ACT test scores and those that do not.

Figure A10.3

The screenshot shows the Denison University website's Test Optional Policy page. At the top left is the Denison logo in a red box. To the right are navigation links: Map / Tour, Events, Give, and an Apply button. Below the navigation is a breadcrumb trail: Home / Forms, Policies, Publications / Test Optional Policy. The main heading is "Test Optional Policy". Underneath, the "Admission" section states: "Denison practices test-optional admission, meaning that applicants are not required to submit standardized test scores as part of their application." A horizontal separator line follows. The "Overview" section states: "If applicants choose this option, we will place additional weight on the rigor of their high school curriculum and their performance in the courses they completed. If applicants desire, they may provide ACT, SAT, and/or SAT Subject Test scores as additional information in support of their application for admission."

Notes: This figure is a screenshot of the admissions page for a test-optional school, Denison University. They indicate that applicants withholding their standardized test scores will be scrutinized more heavily by the rigor of their high school coursework, as well as their performance in their respective classes.

Appendix B

Additional Material for Chapter 2

B.1 Survey Design

B.1.1 Instructions and Questions

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

[Overall Instructions]

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 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].

[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.

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 [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.

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].

[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.

- 18-24
- 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 Degrees (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.”

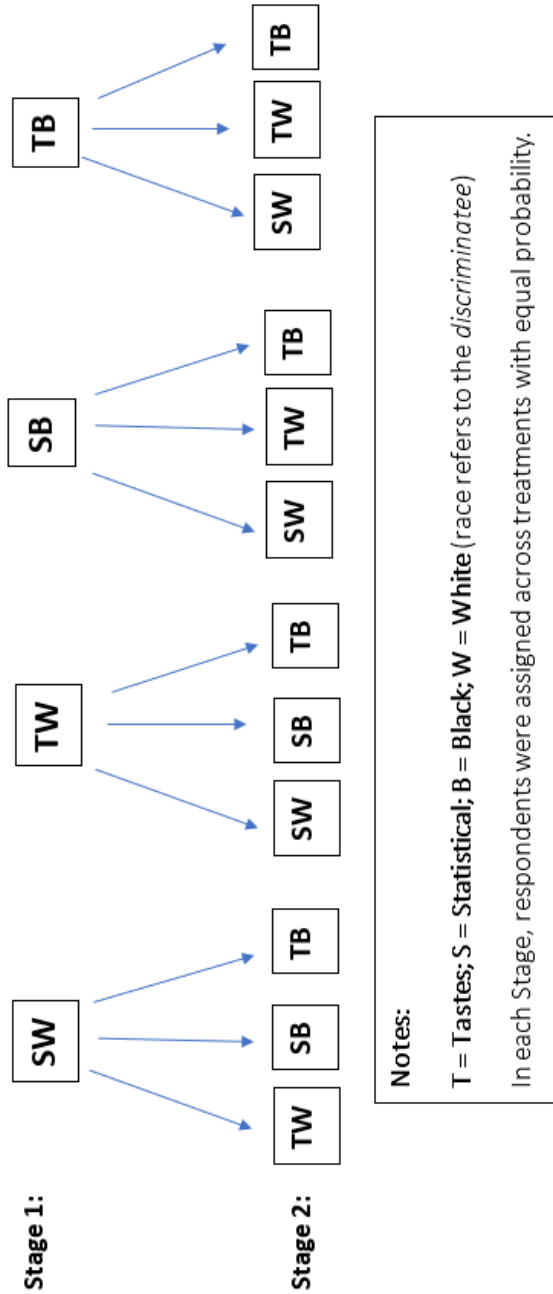
Please do not exit the survey from this page. You must click on the “next button” to reach the “end of survey” page so that your responses are recorded. This button will appear in a few seconds.

B.1.2 Randomization

As illustrated in Figure B1.2.1, subjects were randomly assigned to one of four treatment combinations in Stage 1 of the Survey. In Stage 2, subjects were randomly re-assigned to one of the three treatment combinations they had not encountered in Stage 1. Within each Stage, the more- versus less-justifiable versions of the scenarios for that treatment combination were administered in random order.

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

Figure B1.2.1: Randomization in Stages 1 and 2



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.

B.2 Representativeness

Table B2.1 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 B2.2 shows the mean shares of respondents by political orientation of our MTurk respondents in column (1). Column (2) contains these means from the General Social Survey (GSS), another nationally representative survey sample.¹ Overall, Table B2.2 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 B2.3 and B2.4 compare the geographical distribution of our MTurk sample obtained from the approximate geocoordinates of respondents recorded by the survey

¹Since 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 B2.2 for details).

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 B2.1 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.

Table B2.1: Demographic Composition of MTurk Sample versus the American Community Survey (ACS)

CHARACTERISTIC	MTurk Sample (1)	2019 ACS Sample (2)
Male	0.600	0.4850
Female	0.400	0.5150
White respondent	0.780	0.713
Black respondent	0.115	0.090
Asian respondent	0.042	0.084
Hispanic respondent	0.037	0.020
Indigenous respondent	0.009	0.010
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	0.261
High School or less	0.098	0.362
2-year or some college	0.196	0.307
4-year college or university	0.519	0.203
Higher degree	0.187	0.128
Observations	642	846,557

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

Table B2.2: Composition of MTurk Sample versus the General Social Survey (GSS), by Political Leaning

CHARACTERISTIC	MTurk Sample of (1)	GSS Sample (2)
Extremely conservative	0.101	0.051
Conservative	0.164	0.168
Sightly 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 B2.3: Composition of MTurk Sample by Census Region

CENSUS REGION	MTurk Sample of (1)	2019 ACS Sample (2)
Northeast	0.238	0.178
New England	0.028	0.048
Middle Atlantic	0.210	0.130
Midwest	0.189	0.212
East North Central	0.136	0.146
West North Central	0.053	0.066
South	0.394	0.376
South Atlantic	0.251	0.201
East South Central	0.047	0.059
West South Central	0.097	0.116
West	0.179	0.234
Mountain	0.051	0.073
Pacific	0.128	0.161
Observations	642	1,776

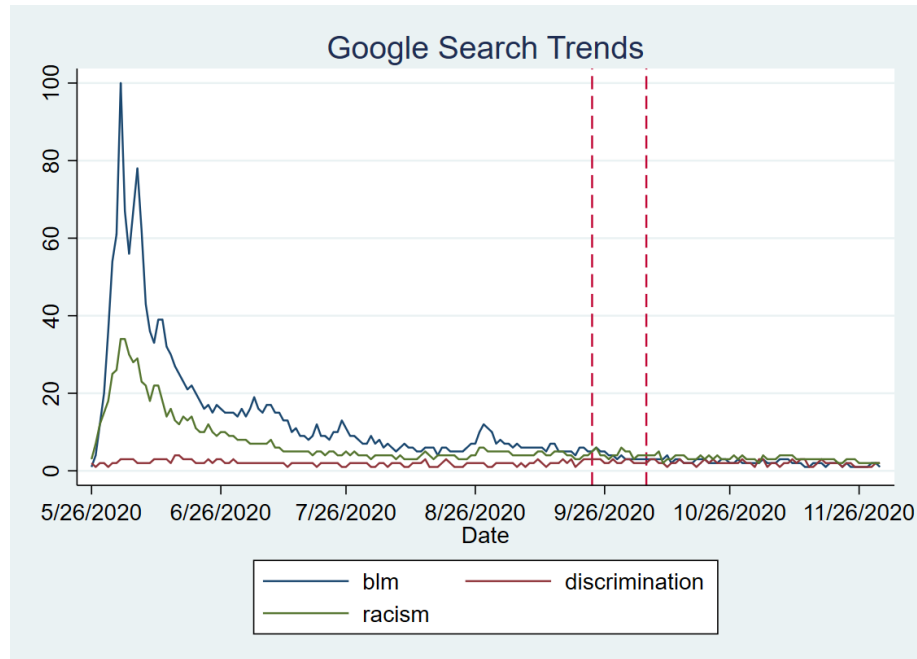
Notes: Column 1 contains the percentage of respondents across U.S. census regions and their respective divisions. Column 2 contains these percentages for the 2019 American Community Survey (ACS) sample for comparison.

Table B2.4: 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 B2.1: Frequency of Google Searches for BLM and Related Keywords around the Survey Date



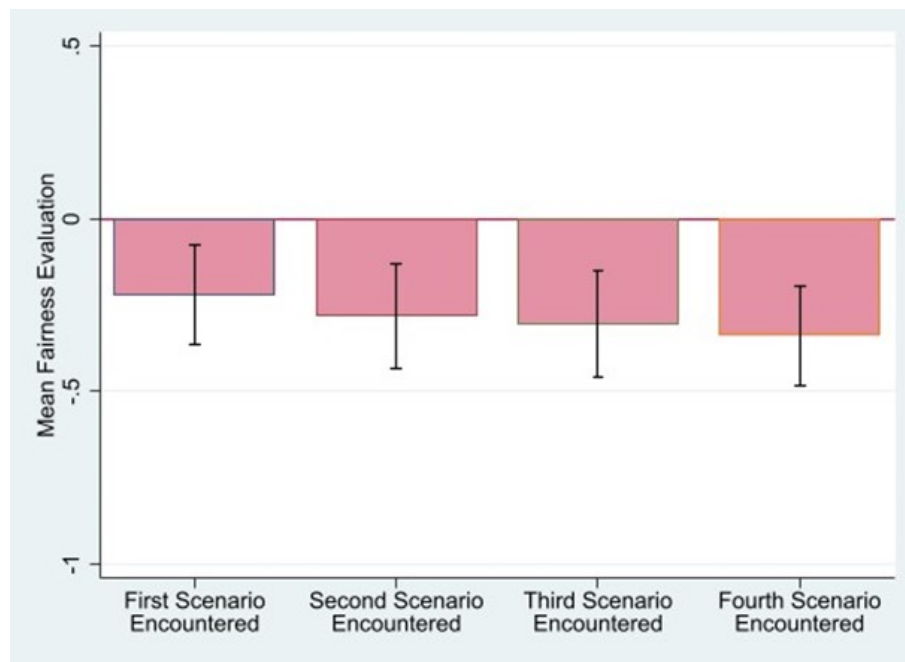
Notes: This figure illustrates trends in Google searches for keywords related to three topics: “Black Lives Matter (blm), racism, and discrimination. The vertical axis represents search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity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

B.3 Order Effects

B.3.1 Pure Order Effects

Figure B3.1.1 shows there is no strong association between the respondents' fairness evaluations and the order of scenarios they encountered throughout the survey.

Figure B3.1.1



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

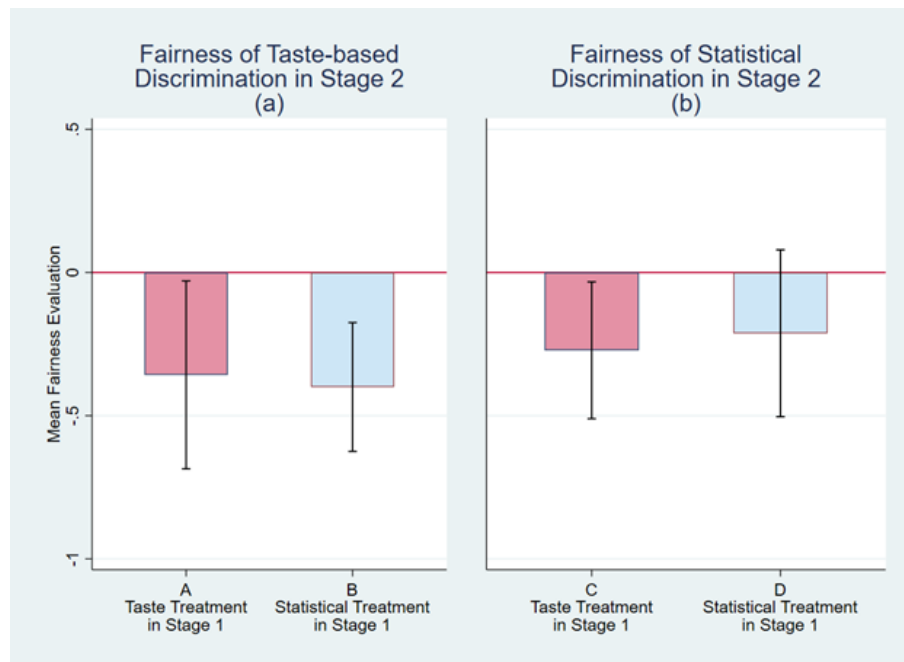
B.3.2 Order Effects for the Taste versus Statistical Treatments

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.

B.3.2.1 Stage 2 Assessments as a Function of Stage 1 Treatment

Figure B3.2.1 (a) shows that Respondents who encountered Taste-based scenarios in Stage 1 view Statistical and Taste discrimination as equally fair in Stage 2. Figure B3.2.1 (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 B3.2.1: Stage 2 Fairness Assessments by Stage 1 Treatment – Taste versus Statistical



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

- A vs B = 0.834
- C vs D = 0.755
- A vs C = 0.675
- B vs D = 0.314

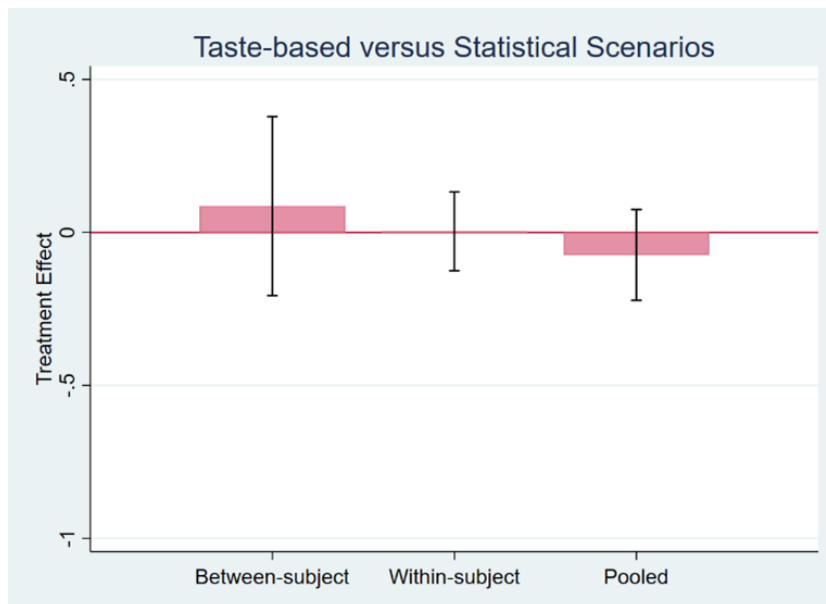
B.3.2.2 Ratings Changes of Subjects Who Switched Treatments

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).

B.3.2.3 Comparing within-subject, between-subject and pooled estimates of the Taste treatment effect

Figure B3.2.3 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 B3.2.3: Comparison of Taste Treatment Effect Estimates



Notes: 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

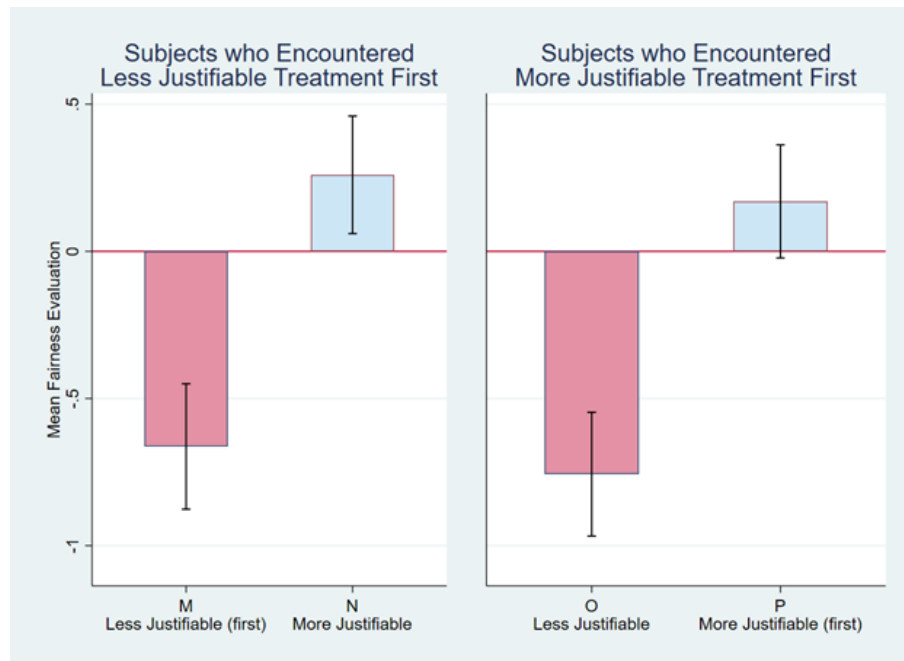
B.3.3 Order Effects for the Less versus More *Justifiable* Treatments

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.

B.3.3.1 *Justifiability* Treatment Variation within Stage 1

Figure B3.3.1 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 B3.3.1: Mean Fairness Ratings by the First Scenario Encountered in Stage 1



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

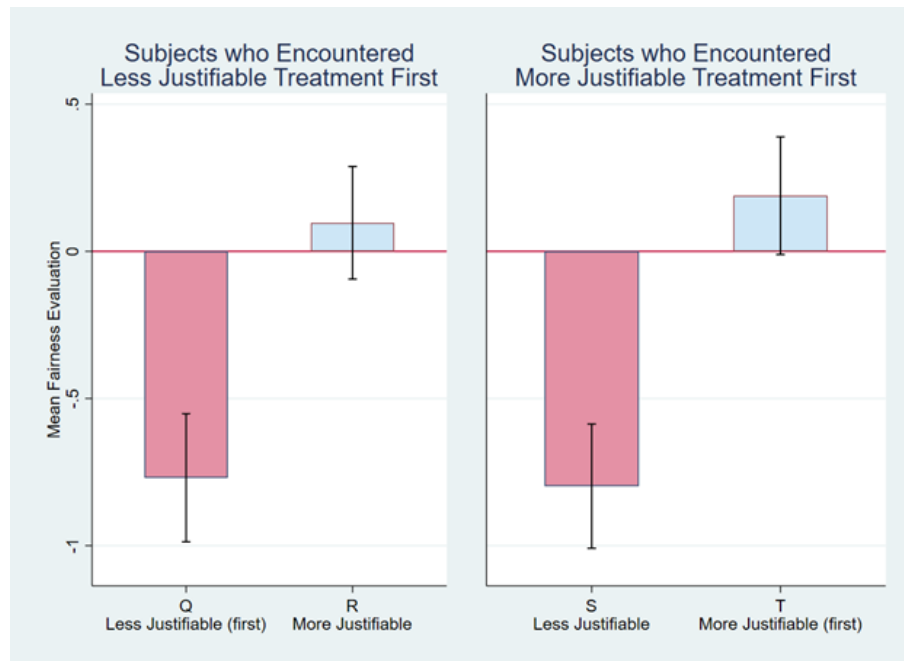
- M vs. N = 0.000
- O vs. P = 0.000
- M vs. O = 0.537
- N vs. P = 0.521

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

B.3.3.2 *Justifiability* Treatment Variation within Stage 2

Figure B3.3.2 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 B3.3.2: Mean Fairness of Respondents by the First Scenario they Encountered in Stage 2



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

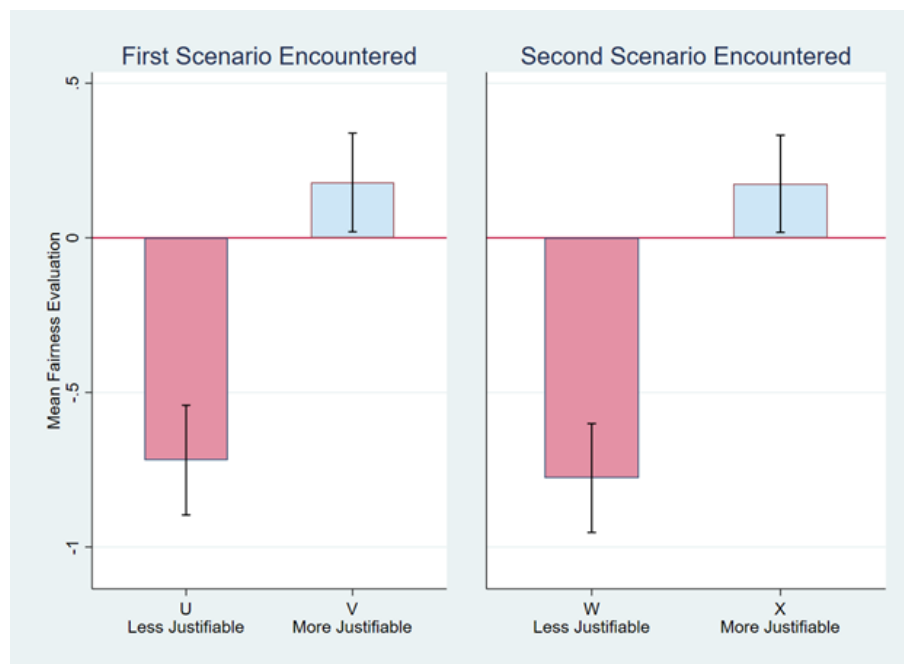
- Q vs. R = 0.000
- R vs. S = 0.000
- Q vs. S = 0.854
- R vs. T = 0.513

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

B.3.3.3 Pooling within-Stage *Justifiability* Treatment Variation from both Stages

Figure B3.3.3 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 B3.3.3: Mean Fairness of Respondents by the First Scenario they Encountered in Stage 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: $U - V = W - X$: $p = .782$

B.3.3.4 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 B3.3.4 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 B3.3.4: Mean Fairness of Respondents by the First Scenario they Encountered in Stage 2



Notes: 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

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.

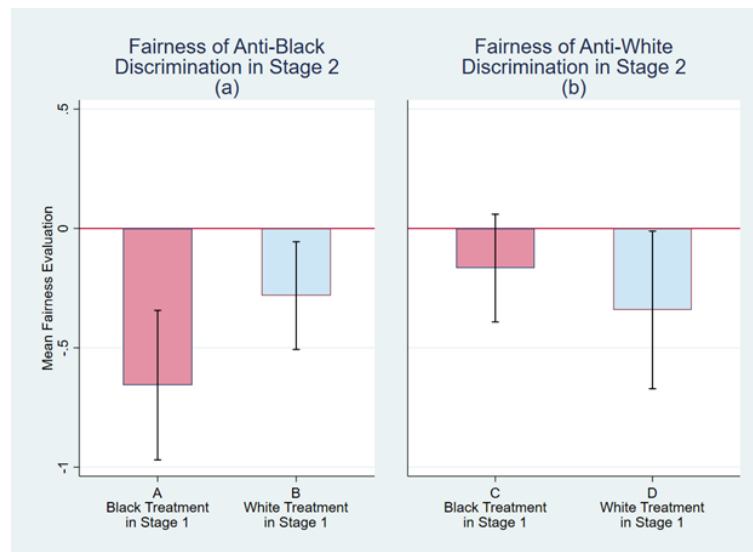
B.3.4 Order Effects for the Race Treatment

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).

B.3.4.1 Stage 2 Assessments as a Function of Stage 1 Treatment

Figure B3.4.1 (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 B3.4.1: Mean Fairness of Respondents by the First Scenario they Encountered in Stage 2



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

- A vs B = 0.055
- C vs D = 0.385
- A vs C = 0.012
- B vs D = 0.767

B.3.4.2 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 ($p = .000$).

B.3.4.3 Comparing within-subject, between-subject and pooled estimates of the Race treatment effect

Figure B3.4.3 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 B3.4.3: 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

B.4 Exploring the Effects of Education on Fairness Ratings

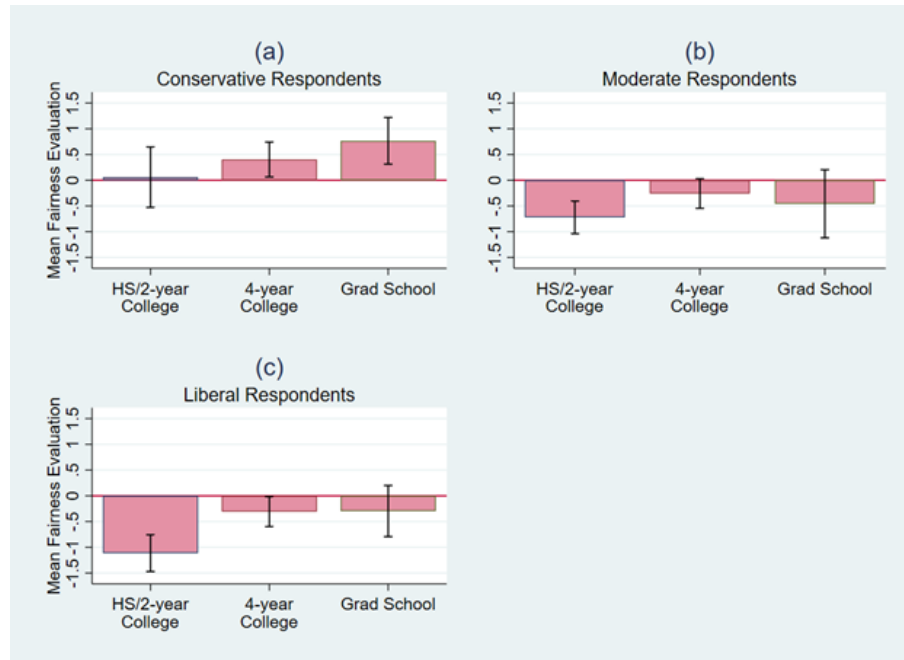
This Section explores the unexpected (to us) positive association between respondents' education and their ratings of the fairness of discriminatory actions. We show, first of all, that the positive association between education and fairness is not an artifact of political differences between the education groups. Instead, Figure B4.1 shows that education is associated with increased perceived fairness within each of our three political groups. Next, while our respondents' political leanings affect the way they respond to our Race treatment, we show that education does not have this effect: Despite being more tolerant of discriminatory acts in general, respondents of all education levels react more negatively to anti-Black and to anti-White discrimination (Figure B4.2). In fact, this discriminatee race effect is remarkably constant across education groups, despite the differences in their mean fairness assessments.

Finally, one of our main findings in the paper is that conservatives do not exhibit a discriminatee race effect, while moderates and liberals do. In Figure B4.3, we show that education differences do not account for this fact either. In fact, our that liberals exhibit a discriminatee race effect and conservatives do not is present within all three education groups (Figure B4.3).

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

Figure B4.1: Mean Fairness Assessments by Education and Political Leaning

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

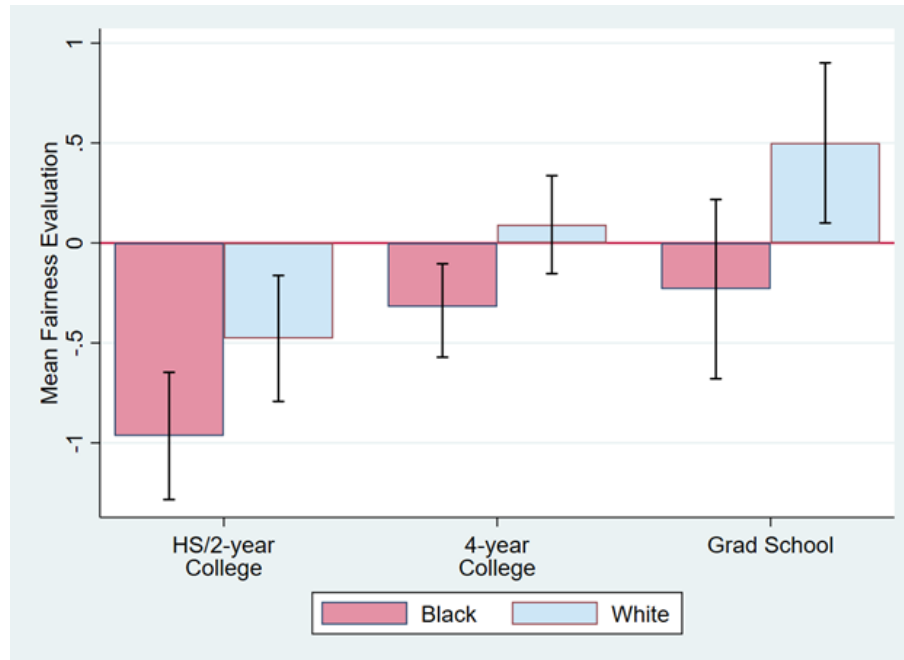


Notes: This figure is based on only Stage 1 observations. The p -values below are clustered by respondent.

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

Figure B4.2: Discriminatee Race Effects by Education

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

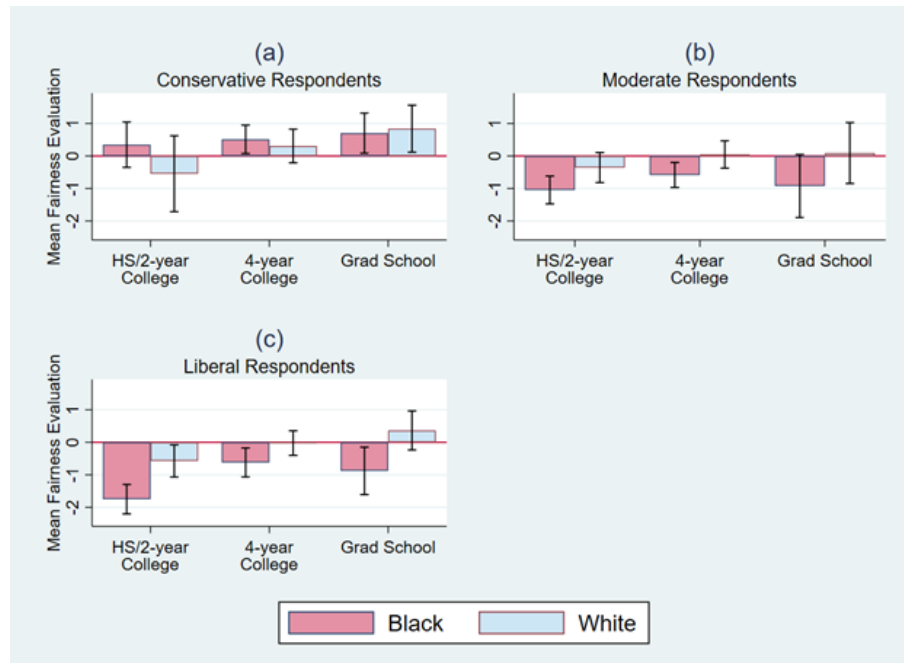


Notes: This figure is based on only Stage 1 observations. The p -values below are clustered by respondent.

- For HS/2-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 B4.3: Discriminatee Race Effects by Education and Political Leaning

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



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

- For conservative respondents:
 - For HS/2-year 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
- For moderate respondents:
 - For HS/2-year 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
- For liberal respondents:
 - For HS/2-year 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

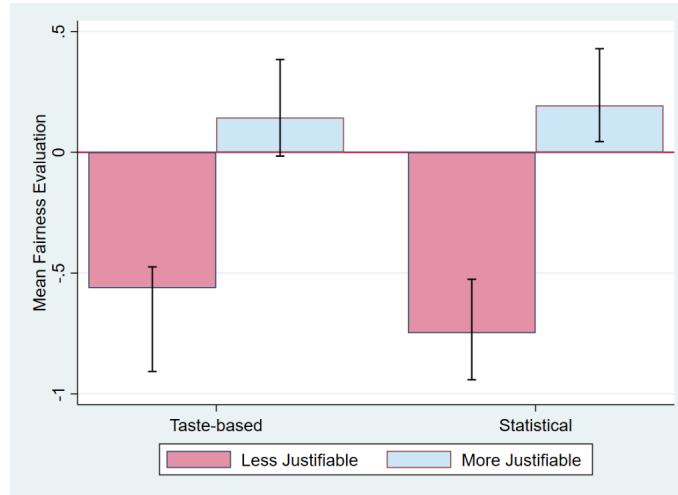
B.5 Robustness Tests for Sections 3 and 4

B.5.1 Replicating Figures 2.2 and 2.3 Using First Scenarios Only

One of our more remarkable findings is that respondents' relative evaluations of the more versus less justifiable scenarios were so similar, regardless of the respondent's political orientation and of the race of the fictitious discriminatee. One might reasonably wonder whether this phenomenon reflects the fact that these two scenario types were always presented after each other and that subjects were asked to pay attention to the differences between the two types. To eliminate the possibility that subjects will be tempted to rank these two scenario types in the same way when they appear in sequence, we now replicate Figure 2.2 of the paper (which was estimated using both scenarios each person saw in Stage 1) using only data from the first scenario each respondent encountered. Remarkably, the results, shown in Figure B5.1.1, are indistinguishable from Figure 2.2. We conclude that subjects' perceptions of the relative fairness of the more- versus less-justifiable scenarios are the same, even when each subject has seen only one of the two scenario types.

Figure B5.1.2 repeats this same exercise for Figure 2.3, which illustrated discriminatee race effects using both scenarios each respondent encountered in Stage 1 of the survey. Figure B5.1.2 shows that the results are extremely similar if we use only information from the very first scenario each respondent encountered in the survey.

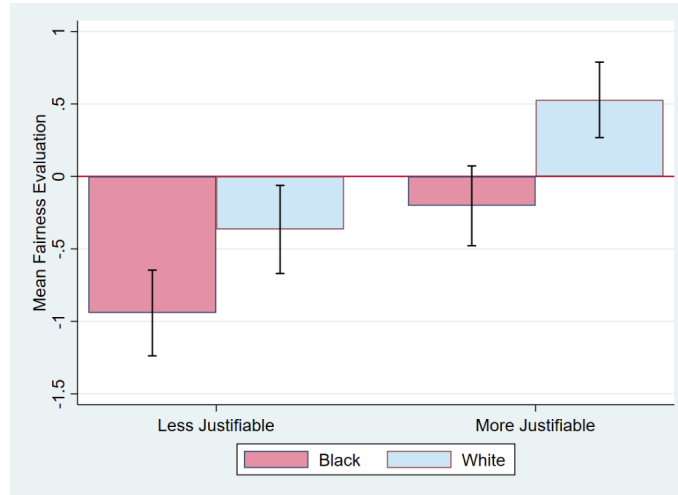
Figure B5.1.1: Fairness Ratings by Type of Discrimination and Justifiability – First Scenario Only



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.

- 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

B5.1.2: Fairness Ratings by Justifiability and Discriminatee Race: First Scenario Only (replicates Figure 3)



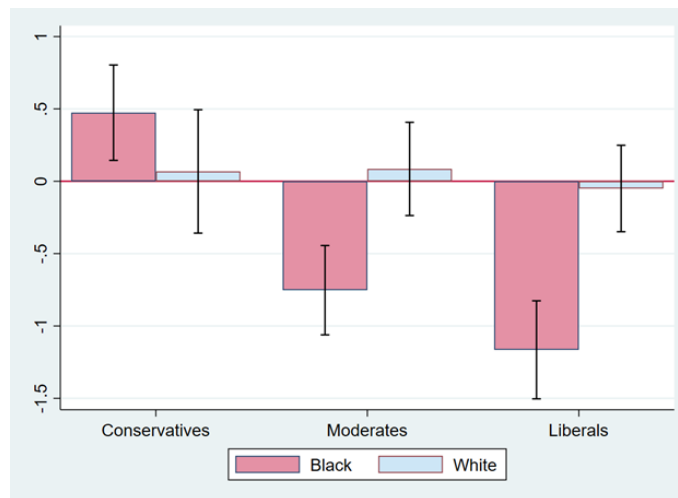
Notes: This figure replicates Figure 3 using only observations from the first scenario encountered by respondents in Stage One. Therefore, the p-values displayed below are not clustered.

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

B.5.2 Discriminatee Race Effects by Political Orientation for White Respondents Only

To probe the in-group bias hypothesis more deeply, here we replicate Figure 2.5 of the paper for White respondents only. The goal is to see if there is evidence of racial in-group bias if we focus on the subset of White respondents who label themselves as conservatives. Interestingly, the discriminatee race effect does switch signs in this group, relative to Figure 2.4 (which includes all respondents): conservative White respondents rate discrimination against Black people as *more* fair than discrimination against White people. This discriminatee race effect is not significantly different from zero at conventional levels, however ($p = .134$).

Figure B5.2.1: Discriminatee Race Effects by Political Orientation, White Respondents Only



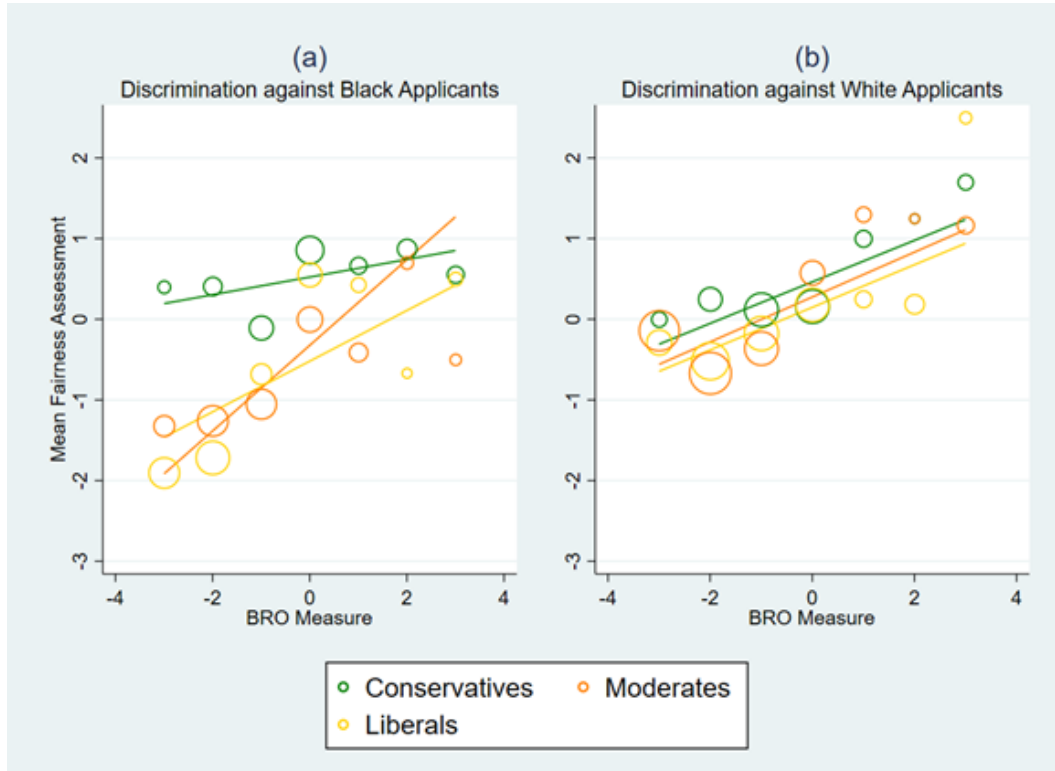
Notes: This figure reproduces Figure 6, but it only reflects the fairness evaluations of White respondents. The p -values below are clustered by respondent.

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

B.5.3 Effects of Perceived Relative Opportunities (BRO) on Fairness Ratings, using Three Political Groups

Figure B5.3 replicates Figure 2.8 of the paper, showing separate results for moderates instead of combining moderates with liberals. For both anti-White and anti-Black discrimination moderates' fairness ratings are quite similar to liberals', and exhibit similar patterns with respect to BRO.

Figure B5.3: Effects of Perceived Relative Opportunities (BRO) on Fairness Ratings, by Discriminatee Race with Three Political Groups



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

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

B.6 Analysis of Open-Text Responses

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

Recall the scenario that you just evaluated, in which [brief description of second scenario encountered in Stage 1].

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

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

Tables B6.1 and B6.2 below summarize the counts of answers in each of these three categories, and provide examples of answers belonging to each category. Among the respondents who said discrimination was unfair or very unfair (Table B6.1), 51 percent (84/166) made a statement to the effect that making a hiring decision based on race was unfair. Another eight percent (14/166) said it was wrong to make a hiring decision on

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

one's tastes. These reasons often overlapped (making it hard to choose which category was most appropriate). Both of them occurred much more often in the tasted-based scenarios. Finally, 41 percent (68/166) said that using statistical information was unfair (e.g. because each individual is different). Essentially all of these answers were for the statistical scenarios; many of them referred to the low quality of information in the less-justifiable statistical scenario. Words like racist, racism, bigoted, discrimination, prejudice, bias, and stereotype were commonly used in all these answers.

Among the respondents who said discrimination was fair or very fair (Table B6.2), missing and hard-to-interpret answers were much more common. With that caveat, 18 of 39 usable answers (46 percent) made a statement to the effect that a business owner's primary responsibility is to ensure their business thrives and survives. Almost all these answers referred to the customer discrimination scenario, where catering to discriminatory customers allowed the employer to 'avoid losing sales). Another 36 percent (14/39) referred to an employer's rights (for example, to hire whomever he wishes, regardless of the reason). Finally, 18 percent (7/39) said that that it was acceptable to make hiring decisions based on statistical information on productivity. All of these responses referred to statistical discrimination scenarios. Notably, however, almost half of them referred to the low-justifiability version, where the hiring decision was based on hearsay. For example, "Well, it was based on some sort of evidence-based reasoning process rather than just a sentiment of not wanting to work with a White person."

Table B6.1: Summary of stated reasons why discrimination was “unfair” or “very unfair”

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

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

Examples of “wrong to use race” statements:

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

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

“Someone’s ability to be hired should never be based off of the color of their skin or opinions of others.”

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

Examples of “wrong to use information” statements:

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

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

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

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

Examples of “wrong to use tastes” statements:

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

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

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

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

Table B6.2: Summary of stated reasons why discrimination was “fair” or “very fair”

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

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

Examples of “business must thrive” statements:

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

“If clients do not like to interact (sic) with white personnel that means that white workers hurt business.”

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

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

Examples of “employer rights” statements:

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

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

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

Examples of “OK to use statistical information”:

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

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

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

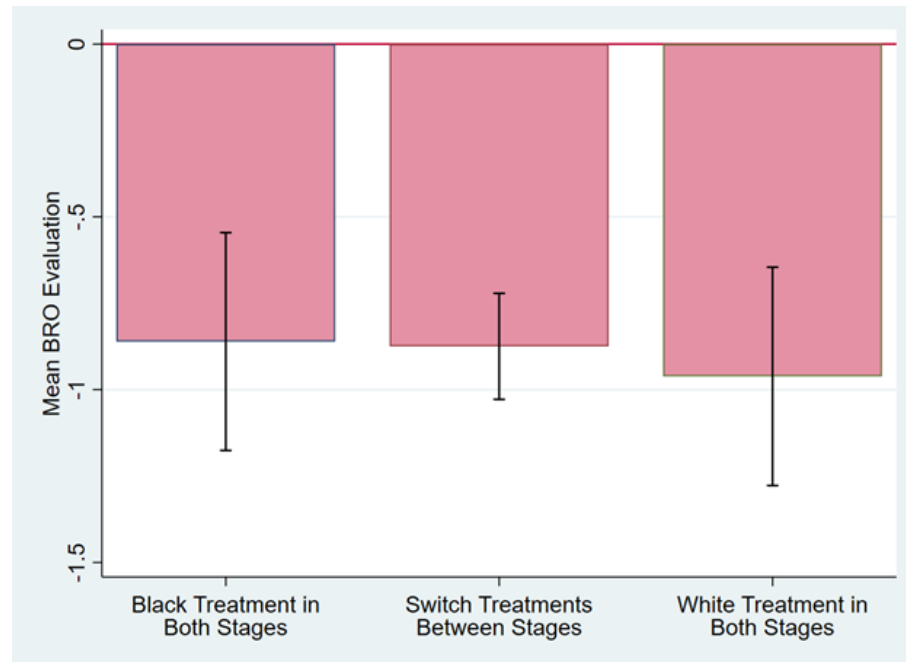
Note: All of these statements (7/7) were for statistical discrimination scenarios.

B.7 Experimenter Demand Effects do not Explain the Race Treatment Order Effects

In Section 2.5.1 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 B7.1 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 B7.2 replicates the analysis for subjects' reported political leaning (on a scale from -3 to +3). Finally, Figure B7.3 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 B7.1: Mean BRO Evaluation Across Respondents' Survey Experience

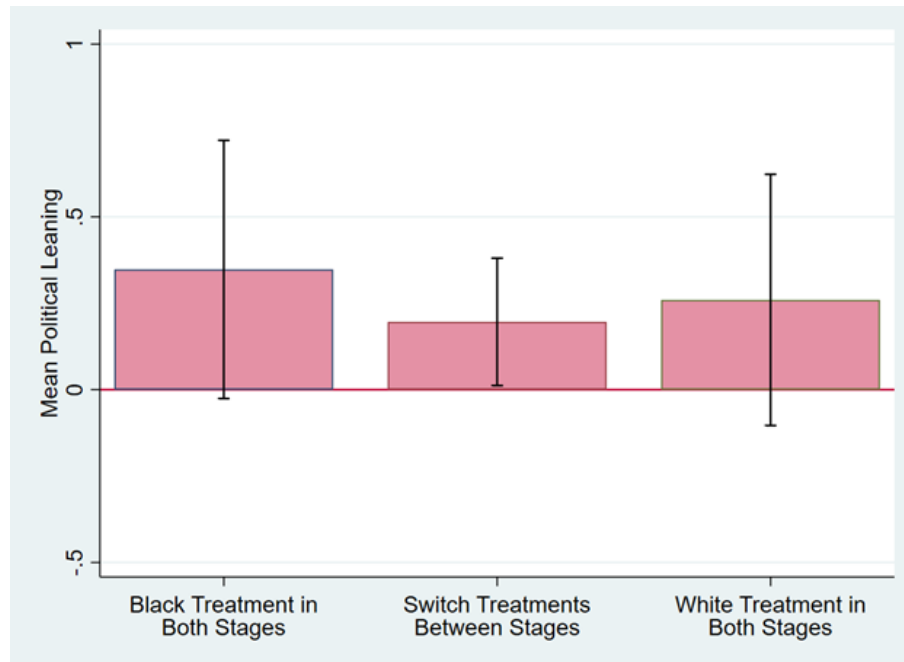


Notes: BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more). The p -values below are clustered by respondent.

- 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 B7.2: Mean Political Leaning Across Respondents' Survey Experience

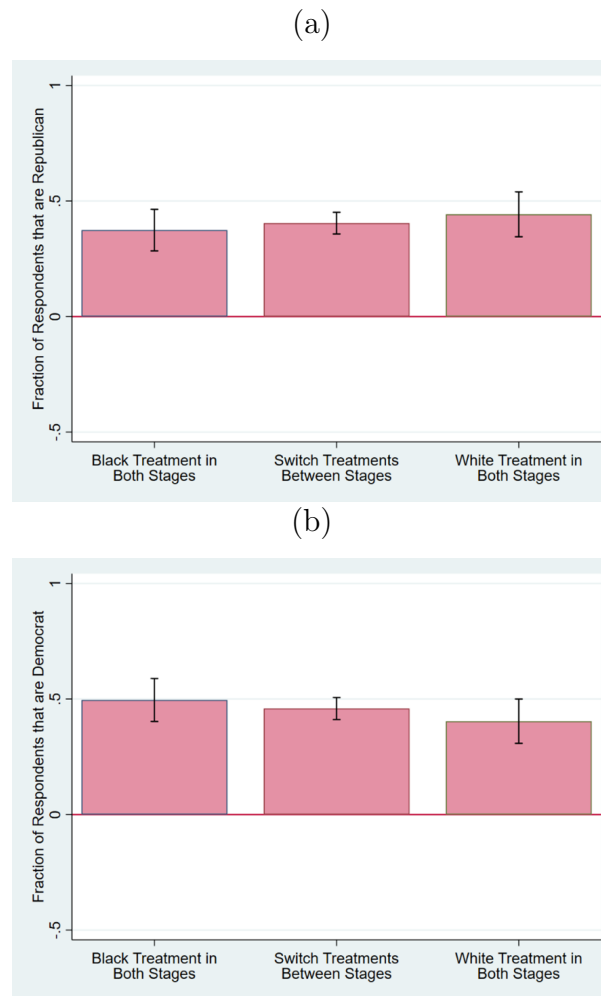


Notes: Political leaning is the respondent's self-description on a scale of -3 (very conservative) to 3 (very liberal). The p -values below are clustered by respondent.

- 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 B7.3 Reported Party Preference Across Respondents' Survey Experience



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

B.8 Estimating α

B.8.1 Splitting the Sample by Groups 1 and 2 (Business Rights Advocates versus Utilitarians)

In this Section, we first document how the race treatment order effect differs between respondent Groups 1 and 2. We show that these order effects are absent in Group 1 (the Business Rights Advocates). In Group 2 (the Utilitarians) the order effects are even stronger than in the aggregate data. We next provide data that allow us to operationalize the ‘trade-off’ model of Group 2’s ratings changes in Section 5.3 of the paper. Specifically, Figures B8.1.1 and Figures B8.1.2 replicate Figure B3.4.1 (which showed that subjects’ Stage 2 fairness assessments depend on the race treatment they encountered in Stage 1) separately for Groups 1 and 2.

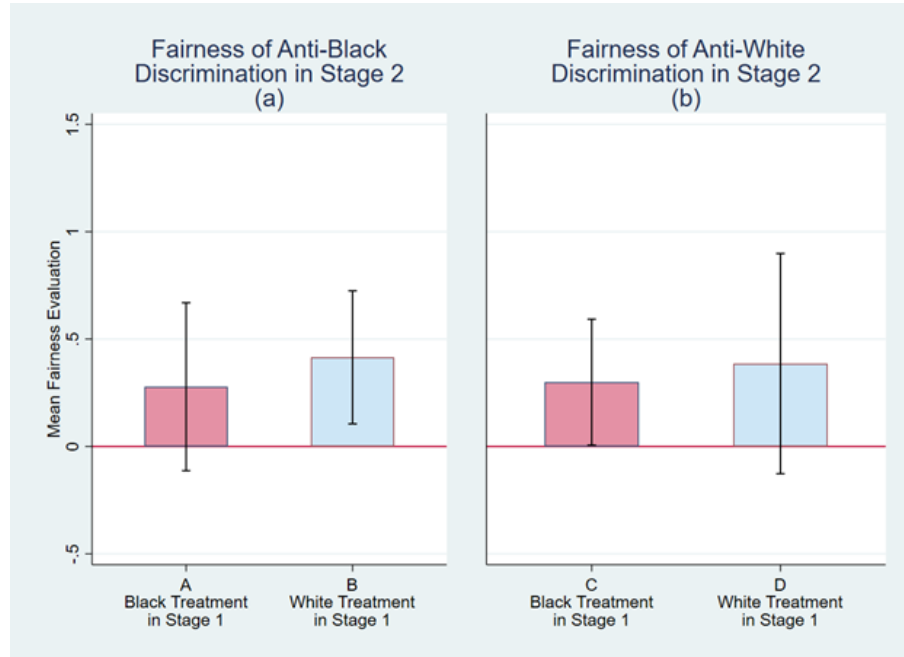
Figure B8.1.1 shows the Stage 2 mean fairness ratings of respondents in Group 1, disaggregated by the race treatments they encountered in both Stages of the experiment. Perhaps the most noteworthy feature is that all the fairness assessments are positive (discrimination is more fair than unfair), but small in value: All the means are between 0 (neither fair nor unfair) and 1 (somewhat fair). Closely related, Group 1’s fairness assessments do not respond to the race treatments, nor do they depend on the order in which the treatments are administered. Specifically, we cannot reject that Group 1’s Stage 2 fairness assessments are unaffected by the treatment they encountered in Stage 1 ($p = .582$ for the Black treatment in Stage 2; $p = .769$ for the White treatment in Stage 2).

Turning to Group 2, Figure B8.1.2 shows a very different pattern. Now all the fairness assessments are negative, but their magnitude is strongly related to the race of the discriminatee. Figure B8.1.2 also shows that Group 2’s Stage 2 fairness ratings do de-

pend on the discriminatee race they encountered in Stage 1. Specifically, respondents who encountered the Black treatment in Stage 2 rated it as much less fair if they also encountered it in Stage 1 than if they encountered a White discriminatee in Stage 1; this difference is highly statistically significant ($p = .013$).

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

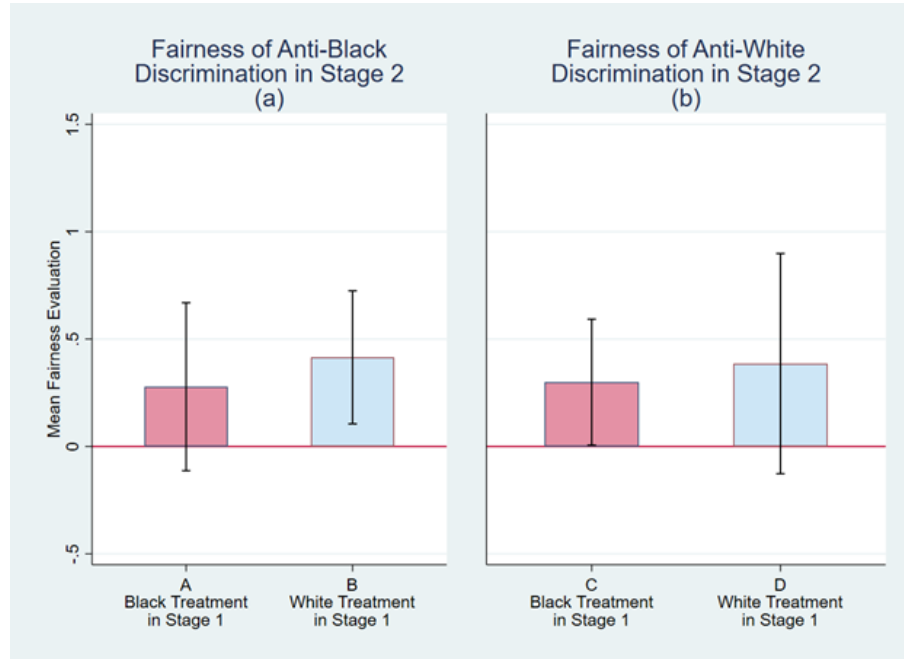
Figure B8.1.1: Race Treatment Order Effects for Group 1 (*Business Rights Advocates*: all conservatives, plus moderates and liberals with $BRO \geq 0$)



Notes: Figure B8.1.1 shows that Group 1's Stage 2 fairness ratings do not depend on the discriminatee race they encountered in Stage 1. The p -values below are clustered by respondent.

- **A vs B = 0.582**
- **C vs D = 0.769**
- A vs C = 0.930
- B vs D = 0.921

Figure B8.1.2: Race Treatment Order Effects for Group 2 (*Utilitarians*: moderates and liberals with $BRO < 0$)



Notes: Figure B8.1.2 shows that Group 2's Stage 2 fairness ratings do depend on the discriminatee race they encountered in Stage 1. Specifically, respondents who encountered the Black treatment in Stage 2 rated it as much less fair if they encountered a Black discriminatee in Stage 1 than if they encountered a White discriminatee in Stage 1. The p -values below are clustered by respondent.

- **A vs B = 0.013**
- **C vs D = 0.269**
- A vs C = 0.002
- B vs D = 0.740

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

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

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

$$B_i^2 = \alpha B_i^* + (1 - \alpha)W_i^1 \quad (\text{B.1})$$

where:

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

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

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

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

Statistically:

- For W-to-B switchers, we can reject both $\alpha = 0$ and $\alpha = 1$ ($p = .000$, $p = .004$).

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

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

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

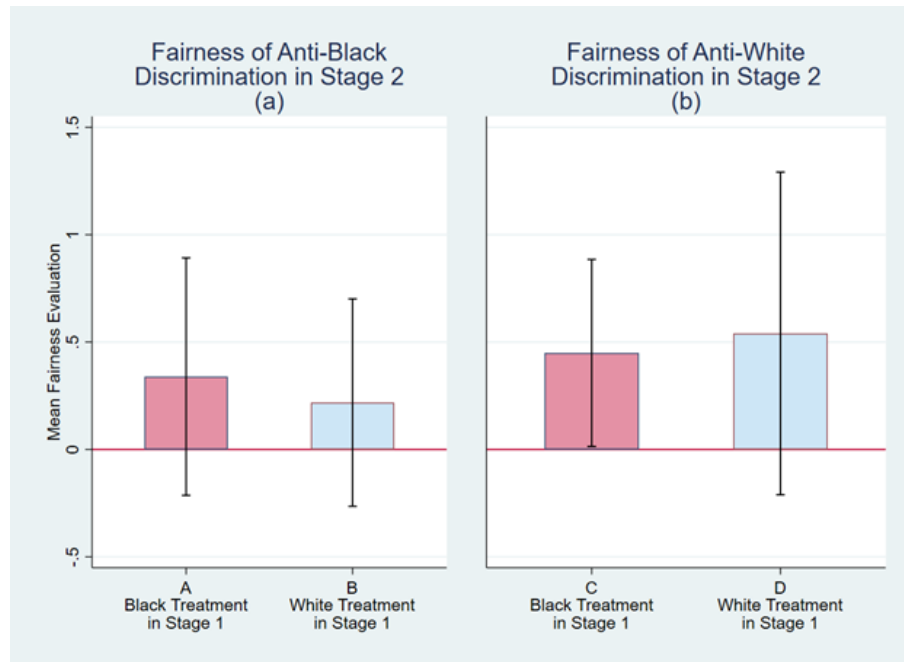
B-to-W Switchers: [0.405, 1.075]

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

B.8.2 Splitting the Sample by Political Leaning (conservatives versus [moderates + liberals])

In this Section, we replicate Appendix B.8.1, splitting the sample by self-reported political affiliation instead of Groups 1 versus 2 (as defined in Section 2.4.4). Since Groups 1 and 2 are predominantly conservative and moderate/liberal respectively, all the results are very similar. Like Group 2, moderates and liberals exhibit a highly significant Race treatment order effect (which we would expect since all Group 2 members are moderate or liberal) and conservatives exhibit no such effect (which we expect since Group 1 is mostly conservative). The estimates of α for [moderates + liberals] are very similar to those for Group 2 as well.

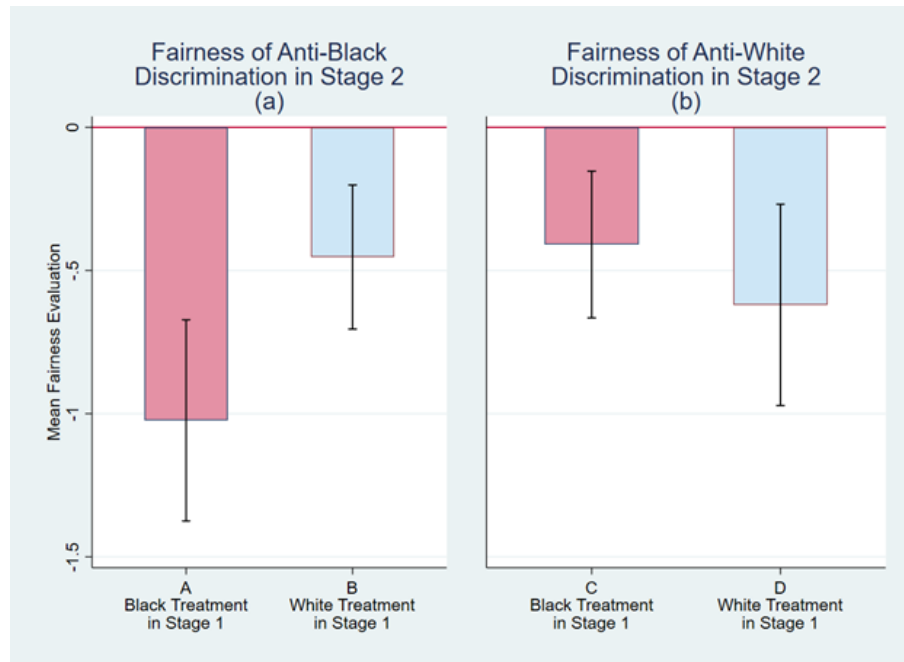
Figure B8.2.1: Race Treatment Order Effects for Conservative Respondents



Notes: Figure B8.2.1 shows that conservative respondents' Stage 2 fairness ratings do not depend on the discriminatee race they encountered in Stage 1. The p -values below are clustered by respondent.

- **A vs B = 0.739**
- **C vs D = 0.829**
- A vs C = 0.750
- B vs D = 0.460

Figure B8.2.2: Race Treatment Order Effects for Conservative Respondents



Notes: Figure B8.2.2 shows that moderate and liberal respondents' Stage 2 fairness ratings do depend on the discriminatee race they encountered in Stage 1. Specifically, respondents who encountered the Black treatment in Stage 2 rated it as much less fair if they also encountered it in Stage 1 than if they encountered a White discriminatee in Stage 1. The p -values below are clustered by respondent.

- **A vs B = 0.009**
- **C vs D = 0.336**
- A vs C = 0.005
- B vs D = 0.443

Using the same method as in Appendix B8.1, we can again calculate α (separately) for W-to-B switchers and B-to-W switchers (among moderate and liberals respondents), yielding:

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

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

Statistically:

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

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

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

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

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

B.9 Replicating the Main Figures with ACS Weights

In this Appendix, we replicate Figures 2.2-2.8 with a set of post-stratification weights. These weights were derived from the 2019 American Community Survey (ACS). They re-weight our MTurk responses by the relative prevalence of our respondents in the ACS in 24 cells, defined by gender (male and female), race (White versus non-White), education (HS/2-year college versus 4-year college or higher) and age (18-24 versus 25-44 versus 45 years of age or older). Table B9.1 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 B9.1 show the sample composition of our MTurk respondents and 2019 ACS respondents at least 18 years old. They show that men and White respondents are modestly over-represented on MTurk. People between the ages of 25 and 44 and four-year college graduates are highly over-represented. Column 2 shows that our weights do quite a good job of correcting for these forms of non-representativeness.

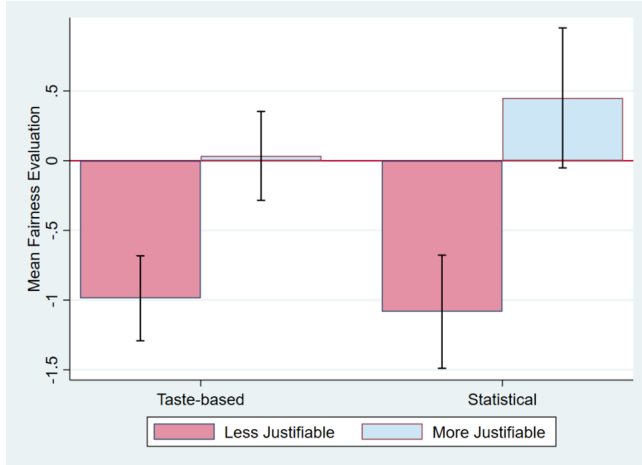
The remaining exhibits in this Appendix replicate Figures 2.2-2.8 using these weights. All the main patterns discussed in the paper are also present here, with one small exception: the weak positive association between BRO and the fairness of discrimination among conservative respondents in Figure 2.8(a) becomes somewhat stronger and statistically significant. Similar to Figure 2.8, however, the slope for conservatives remains much lower than the slope for moderates / liberals.

Table B9.1: 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 respondent	0.780	0.673	0.628
Black respondent	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
High School, 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 B9.1: Fairness Ratings by Type of Discrimination and *Justifiability* (Replicates Figure 2.2)



Less vs More Justifiable Treatment

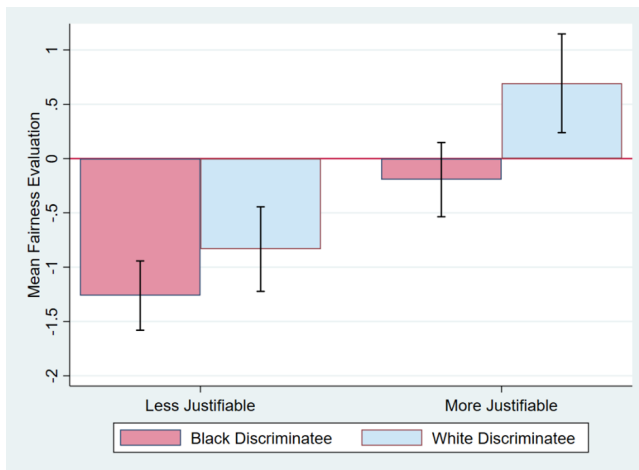
Overall: $p = .000$
 Within taste-based: $p = .000$
 Within statistical: $p = .000$

Taste vs Statistical Discrimination

Overall: $p = .505$
 Within Less-Justifiable: $p = .709$
 Within More-Justifiable: $p = .170$

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

Figure B9.2: Fairness by *Justifiability* and Discriminatee Race (replicates Figure 2.3)



Black vs White Treatment

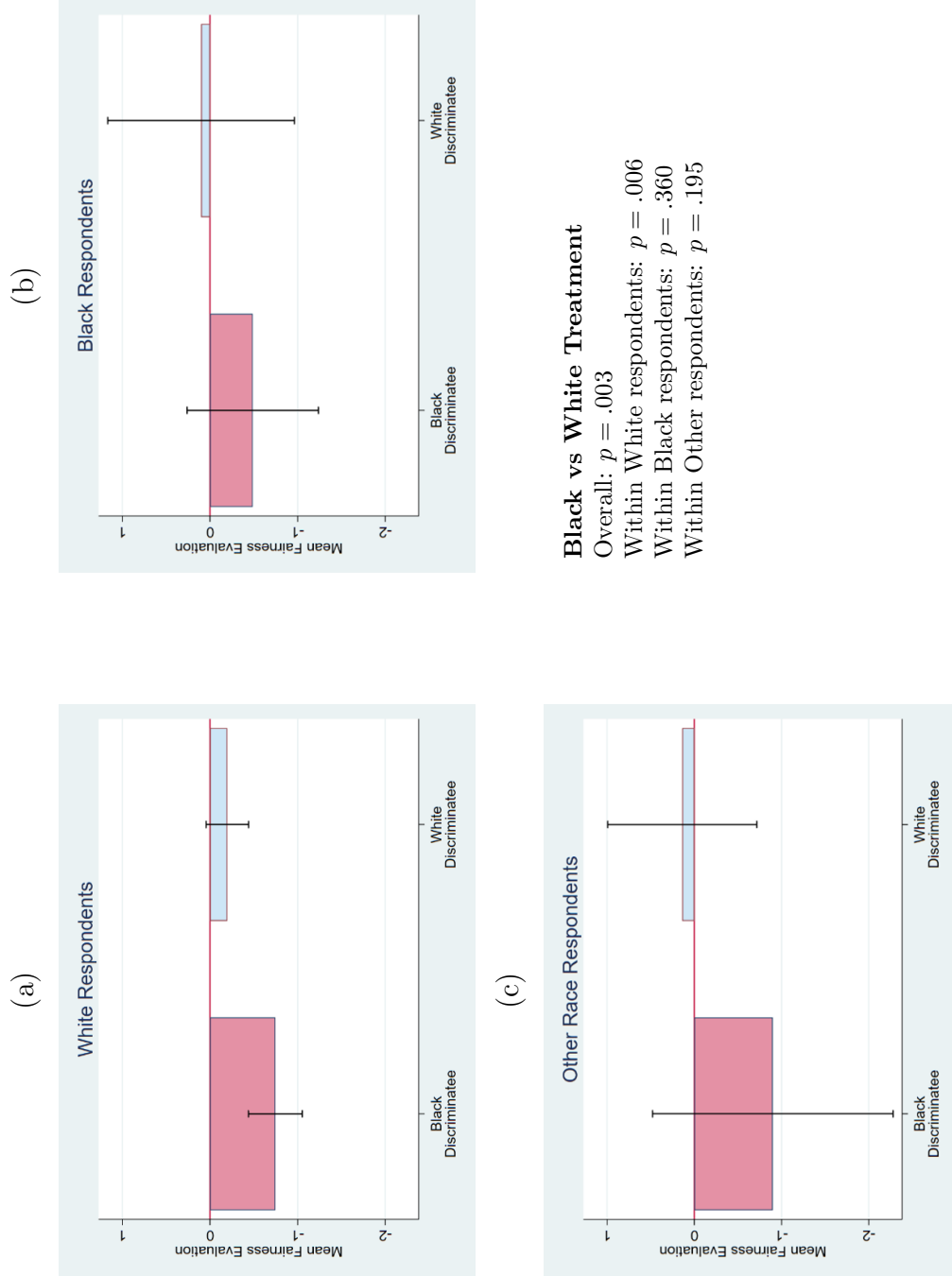
Overall: $p = .003$
 Within taste-based: $p = .095$
 Within statistical: $p = .002$

Less vs More Justifiable Treatment

Overall: $p = .000$
 Within Less-Justifiable: $p = .000$
 Within More-Justifiable: $p = .000$

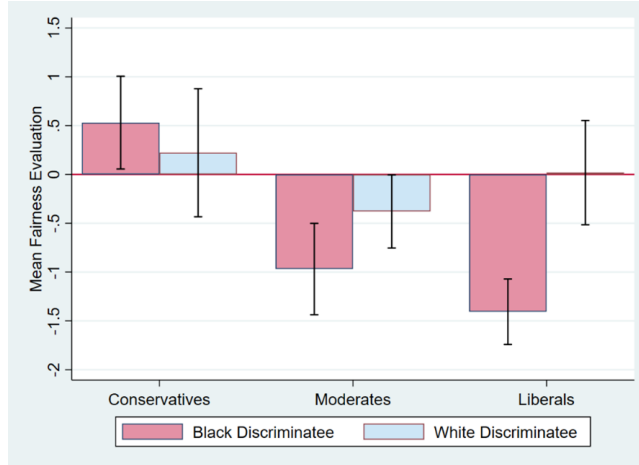
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 1.068 units less fair. Within White Discriminatees, less-justifiable scenarios are 1.527 units less fair. A test for equality of the Less versus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

Figure B9.3 Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 2.4)



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) across all three racial groups yields $p = .832$.

Figure B9.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Figure 2.5)

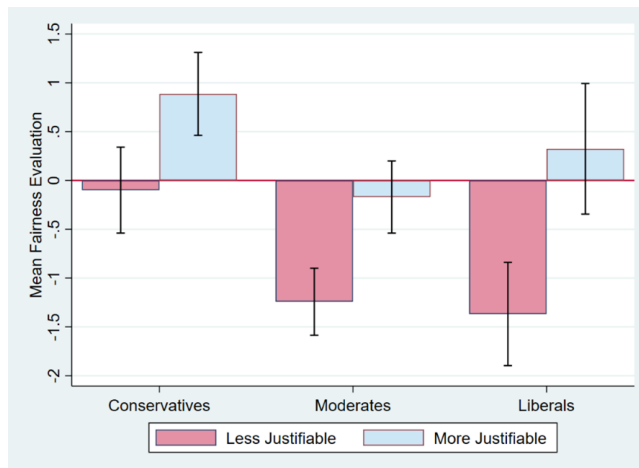


Black vs White Treatment

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 B9.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent’s Political Leaning (replicates Figure 2.6)

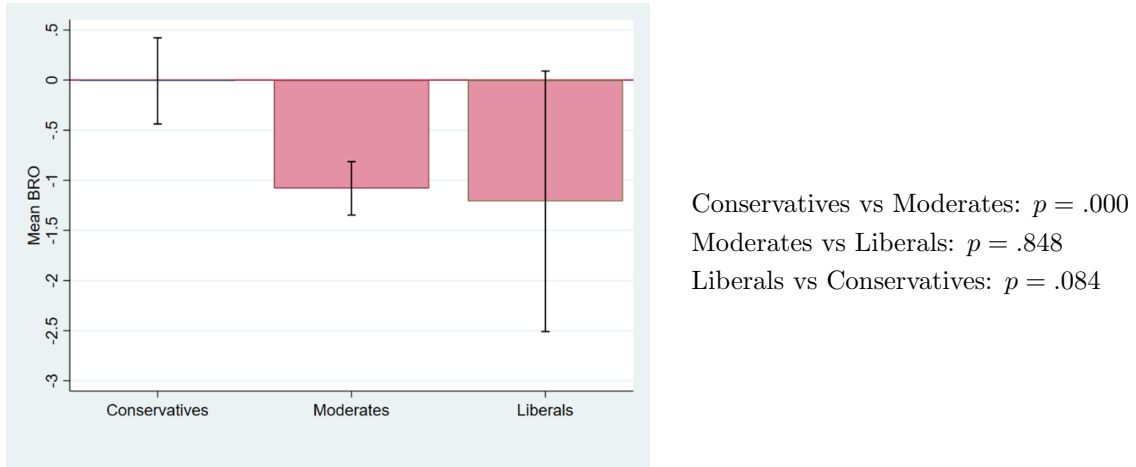


Black vs White Treatment

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

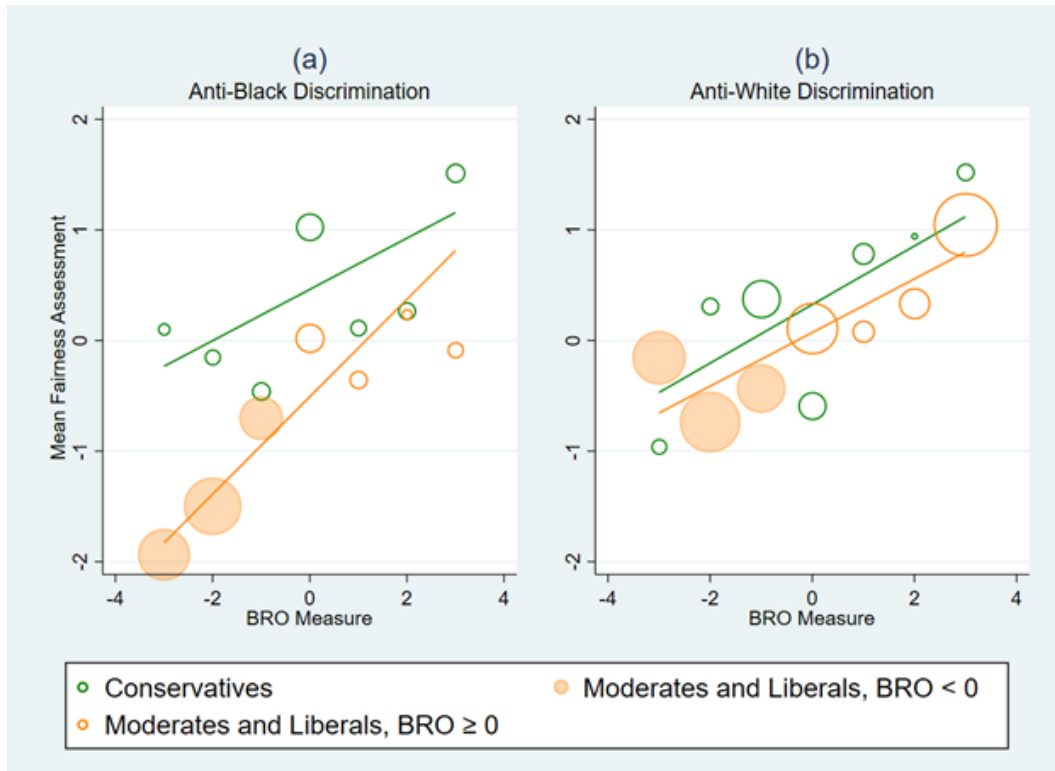
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. A test for equality of the Less versus More Justifiability Gap across Conservatives, Moderates, and Liberals yields $p = .153$.

Figure B9.6 Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 2.7)



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 B9.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 2.8)



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

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

B.10 Replicating the Main Figures with GSS Weights

In this Appendix, we replicate Figures 2.2-2.8 with an alternative set of post-stratification weights. These weights were derived from the 2020 General Social Survey (GSS), and they are based only on a 7-point political leaning scale (i.e., extremely conservative, conservative, slightly conservative, moderate, slightly liberal, liberal, and extremely liberal). Columns 1 and 3 of Table B10.1 show the sample composition of our MTurk respondents and 2020 GSS respondents at least 18 years old. Overall, MTurk respondents differ from the GSS in two main ways: First, compared to the GSS a smaller share of MTurk respondents choose the middle three categories: ‘moderate’ or ‘slightly’ liberal / conservative, while MTurkers are also more likely to locate in the two ‘extreme’ categories. In this sense, MTurkers are politically more extreme than GSS respondents.³ Second, almost identical shares of MTurkers and GSS respondents choose some degree of conservative leaning (ranging from slight to extreme), but many more MTurkers choose some liberal leaning (47.3 versus 30.2 percent). Thus, on average, MTurkers are also more liberal than the U.S. population as a whole. Because our weights do not interact political leaning with any other characteristics, the weighted MTurk sample in column 2 of Table B10.1 mimics the GSS sample perfectly.⁴

The remaining exhibits in this Appendix replicate Figures 2.2-2.8 using these weights. All the main patterns discussed in the paper are also present here. The one exception noted with the ACS weights in Appendix B.9 does not occur here, suggesting that the unusual political mix of MTurkers is not responsible for any of the main results in the paper.

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

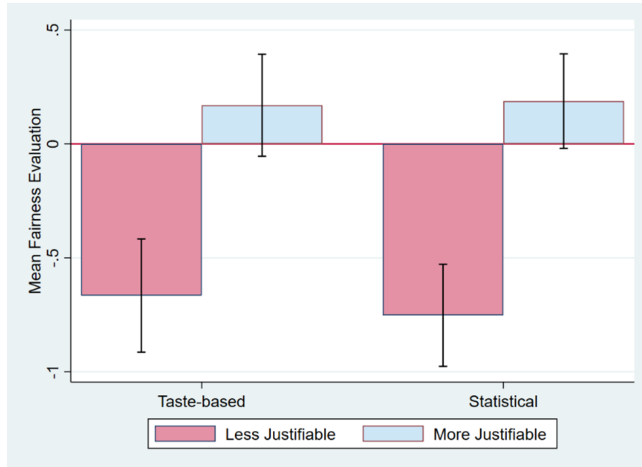
⁴Because of the small size of the MTurk and GSS samples, we did not re-weight our MTurk sample to mimic GSS demographic characteristics; attempts to do this yielded highly extreme and imprecise weights.

Table B10.1: Raw and Re-Weighted Sample composition, GSS weights.

CHARACTERISTIC	MTurk Sample of (1)	Weighted Sample (2)	GSS Sample (3)
Extremely conservative	0.101	0.051	0.051
Conservative	0.164	0.168	0.168
Sightly 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	0.104	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 B10.1: Fairness Ratings by Type of Discrimination and *Justifiability* (Replicates Figure 2.2)



Less vs More Justifiable Treatment

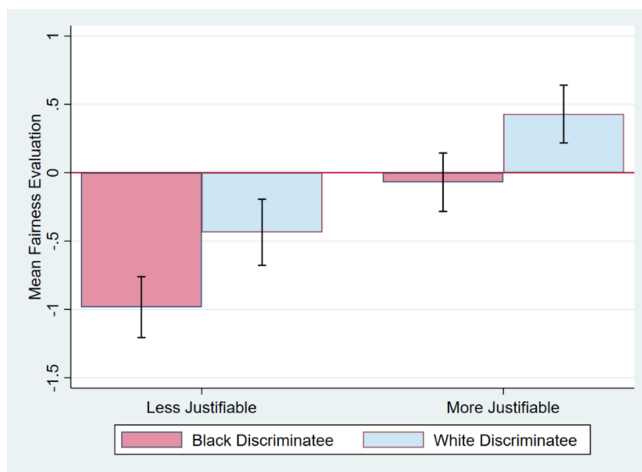
Overall: $p = .000$
 Within taste-based: $p = .000$
 Within statistical: $p = .000$

Taste vs Statistical Discrimination

Overall: $p = .813$
 Within Less-Justifiable: $p = .610$
 Within More-Justifiable: $p = .907$

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

Figure B10.2: Fairness by *Justifiability* and Discriminatee Race (replicates Figure 2.3)



Black vs White Treatment

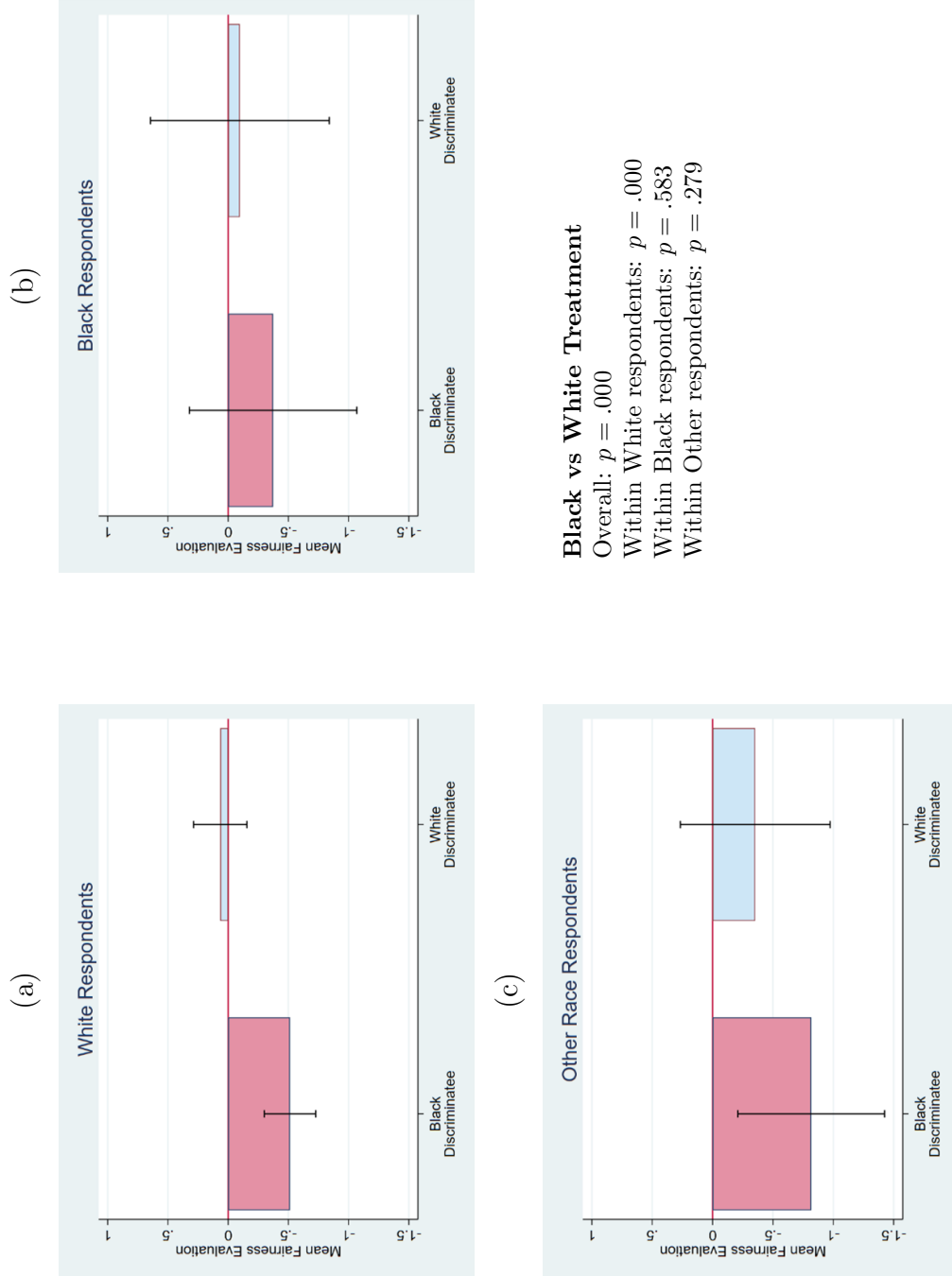
Overall: $p = .000$
 Within taste-based: $p = .001$
 Within statistical: $p = .001$

Less vs More Justifiable Treatment

Overall: $p = .000$
 Within Less-Justifiable: $p = .000$
 Within More-Justifiable: $p = .000$

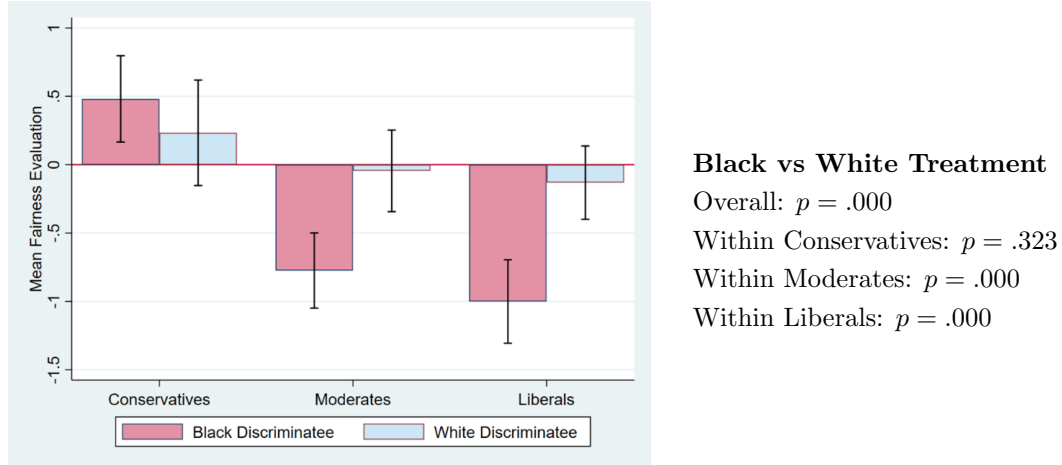
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.914 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.865 units less fair. A test for equality of the Less versus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

Figure B10.3 Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 2.4)



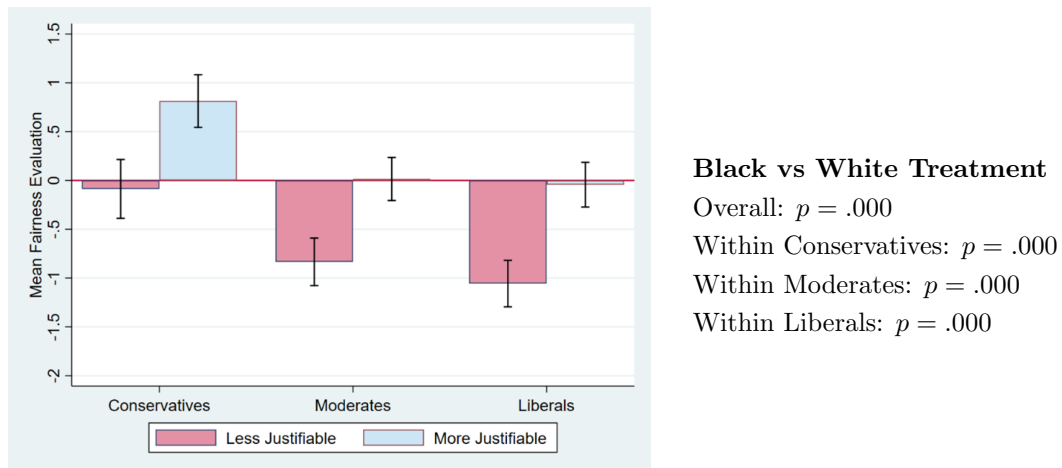
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. A test for equality of the discriminatee race effect (i.e. the Black treatment) across all three racial groups yields $p = .827$.

Figure B10.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Figure 2.5)



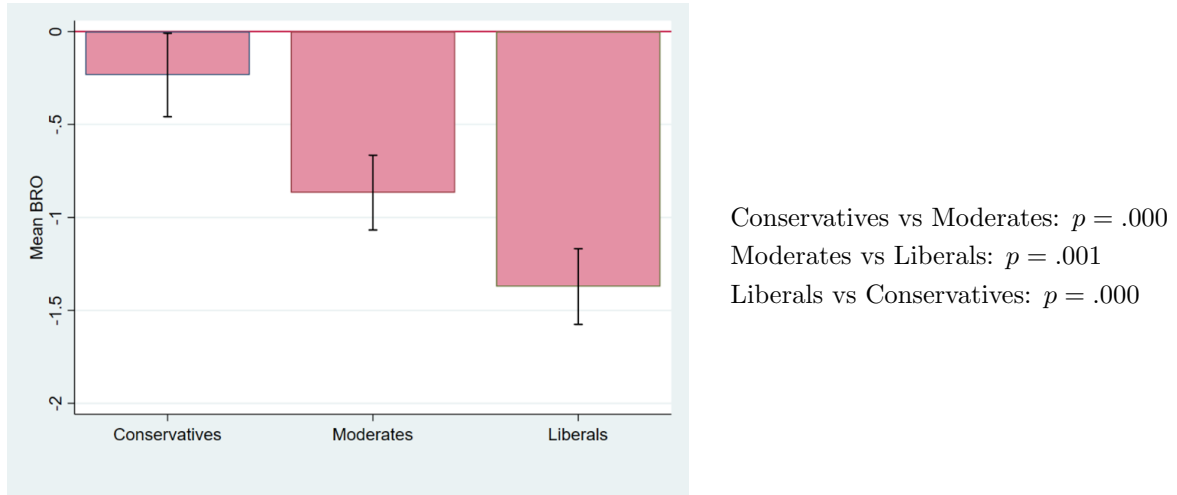
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields $p = .628$. A test for equality between conservatives and (moderates + liberals) yields $p = .001$. s and (moderates + liberals) yields $p = .006$.

Figure B10.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent’s Political Leaning (replicates Figure 2.6)



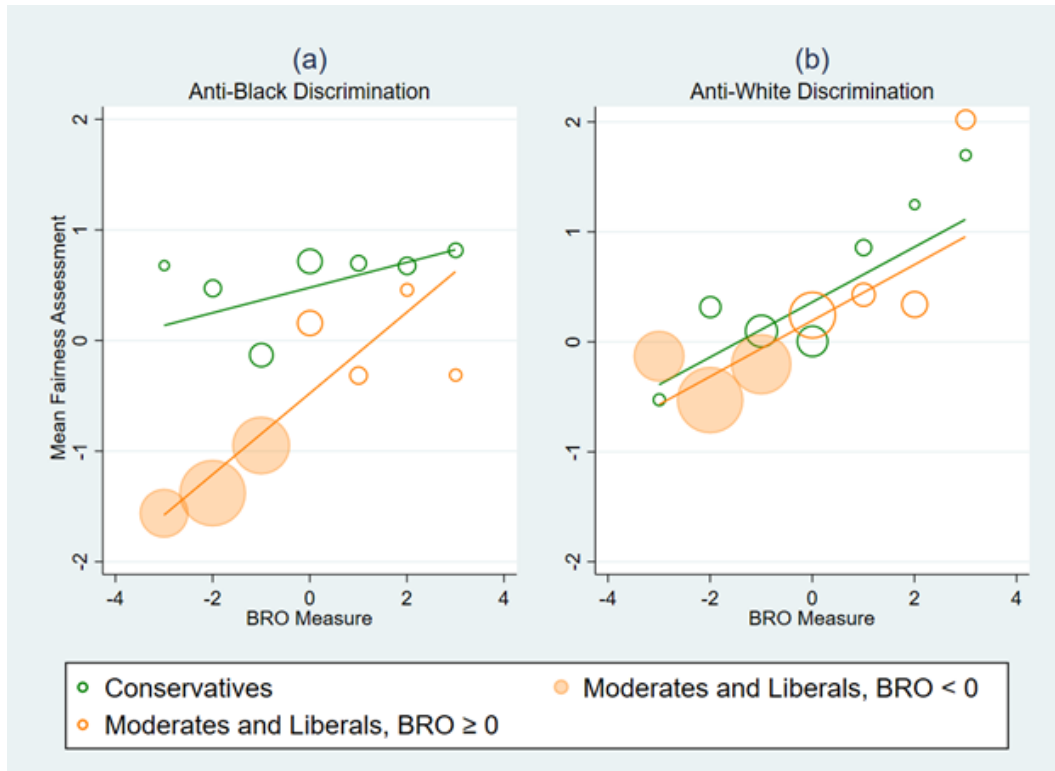
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. A test for equality of the Less versus More Justifiability Gap across Conservatives, Moderates, and Liberals yields $p = .541$.

Figure B10.6 Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 2.7)



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 B10.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 2.8)



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

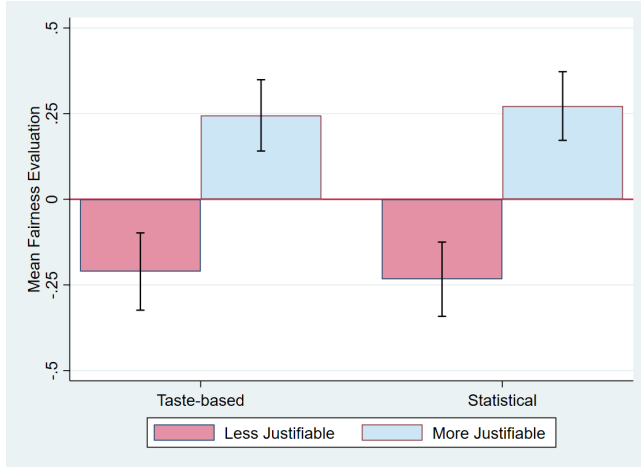
- Panel (a), Discrimination against Black Applicants
 - For Conservatives: slope = 0.114, $p = .231$
 - For Moderates and Liberals, slope = 0.367, $p = .000$
- Panel (b), Discrimination against White Applicants
 - For Conservatives: slope = 0.250, $p = .096$
 - For Moderates and Liberals, slope = 0.254,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

B.11 Replicating the Main Figures with Standardized Fairness Measures

In this section, we replicate the main figures and table by using a standardized version of our fairness ratings. Therefore, all of the means displayed in Figures 2.2-2.8 illustrate deviations from the mean fairness rating for the entire sample, i.e., -0.286 on a scale of -3 to 3 where the standard deviation is 1.920.⁵ We also standardize the BRO (Black relative opportunity) measure, where its mean is -0.886, also on a scale of -3 to 3 where the standard deviation is 1.498. In short, all of the figures are comparable to the ones using the raw fairness and BRO measures.

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

Figure B11.1: Fairness Ratings by Type of Discrimination and *Justifiability* (Replicates Figure 2.2)



Less vs More Justifiable Treatment

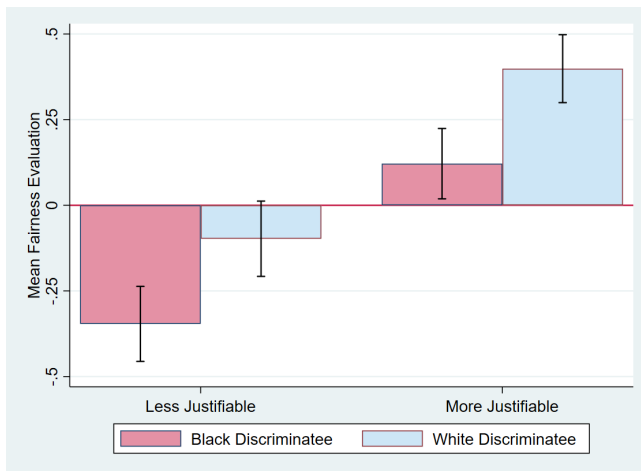
Overall: $p = .000$
 Within taste-based: $p = .000$
 Within statistical: $p = .000$

Taste vs Statistical Discrimination

Overall: $p = .971$
 Within Less-Justifiable: $p = .779$
 Within More-Justifiable: $p = .710$

Notes: This figure is based on Stage 1 observations. 95% confidence intervals are shown. p -values are clustered by respondent.

Figure B11.2: Fairness by *Justifiability* and Discriminatee Race (replicates Figure 2.3)



Black vs White Treatment

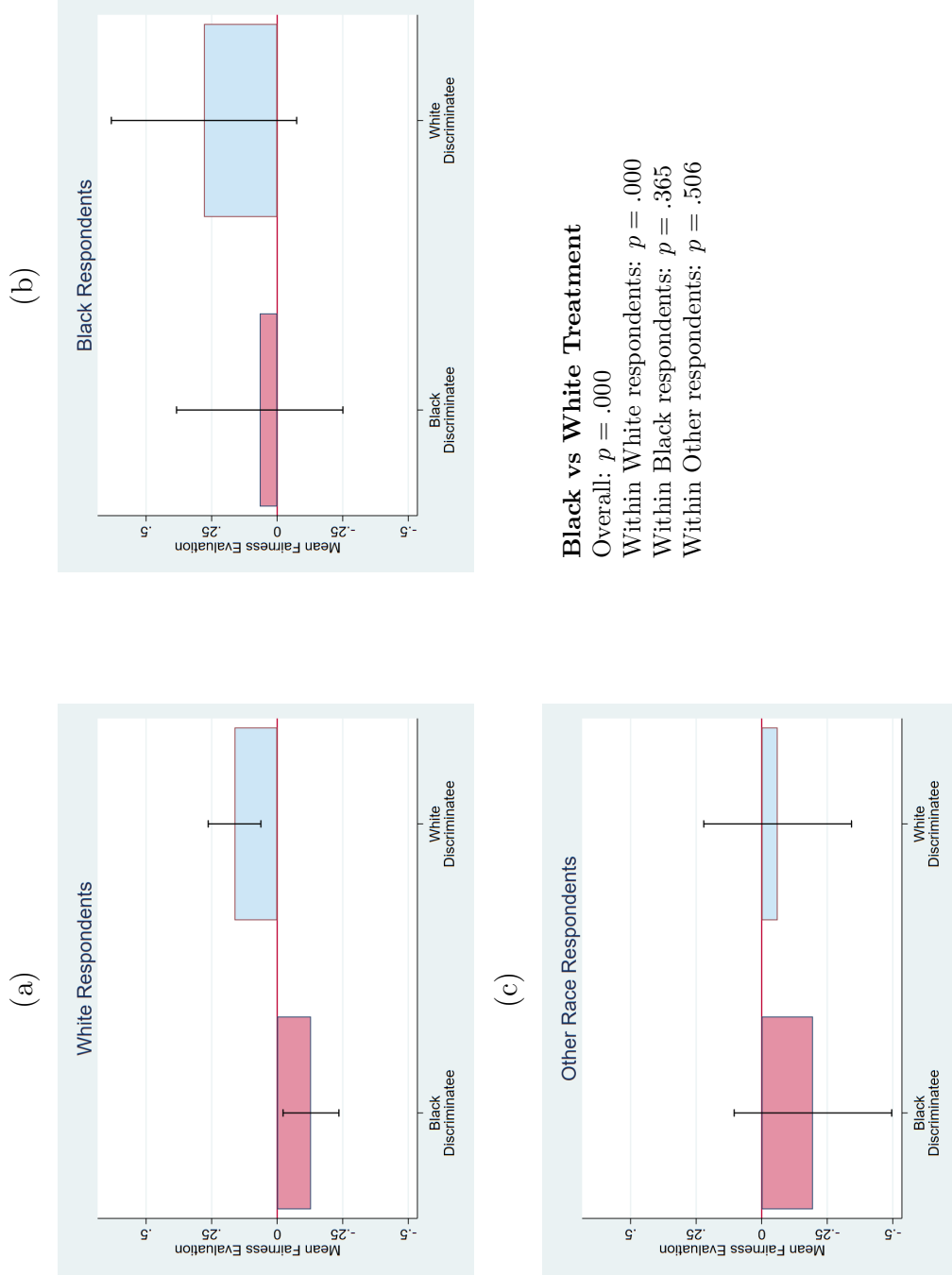
Overall: $p = .000$
 Within taste-based: $p = .002$
 Within statistical: $p = .000$

Less vs More Justifiable Treatment

Overall: $p = .000$
 Within Less-Justifiable: $p = .000$
 Within More-Justifiable: $p = .000$

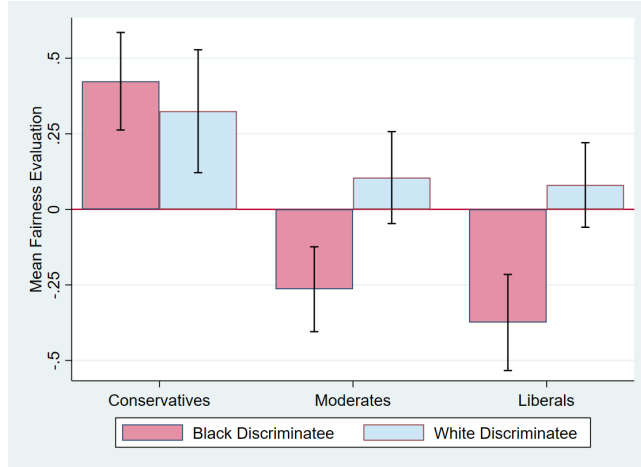
Notes: Figure is based on Stage 1 observations only. 95% confidence This figure is based on only Stage 1 observations. All p -values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.469 standard deviations less fair. Within White Discriminatees, less-justifiable scenarios are 0.495 standard deviations less fair. A test for equality of the Less versus More Justifiability Gap between the Black and White treatment yields $p = .679$.

Figure B11.3 Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 2.4)



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) across all three racial groups yields $p = .739$.

Figure B11.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Figure 2.5)

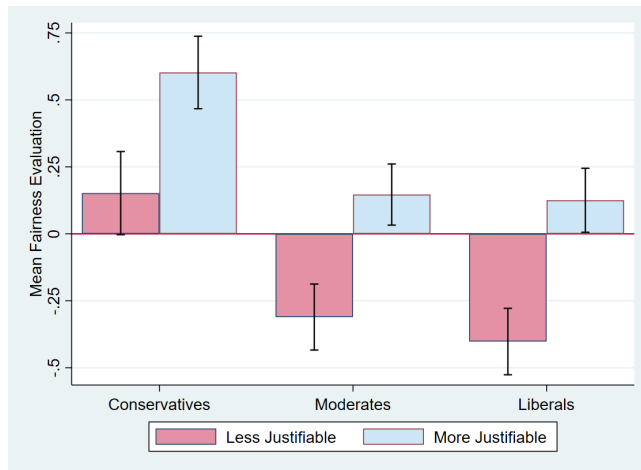


Black vs White Treatment

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 B11.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent’s Political Leaning (replicates Figure 2.6)

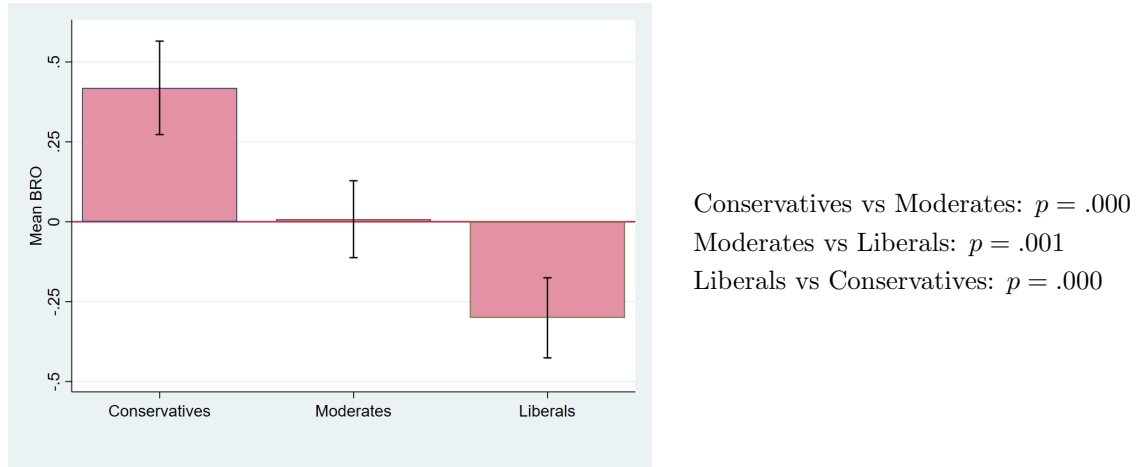


Black vs White Treatment

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

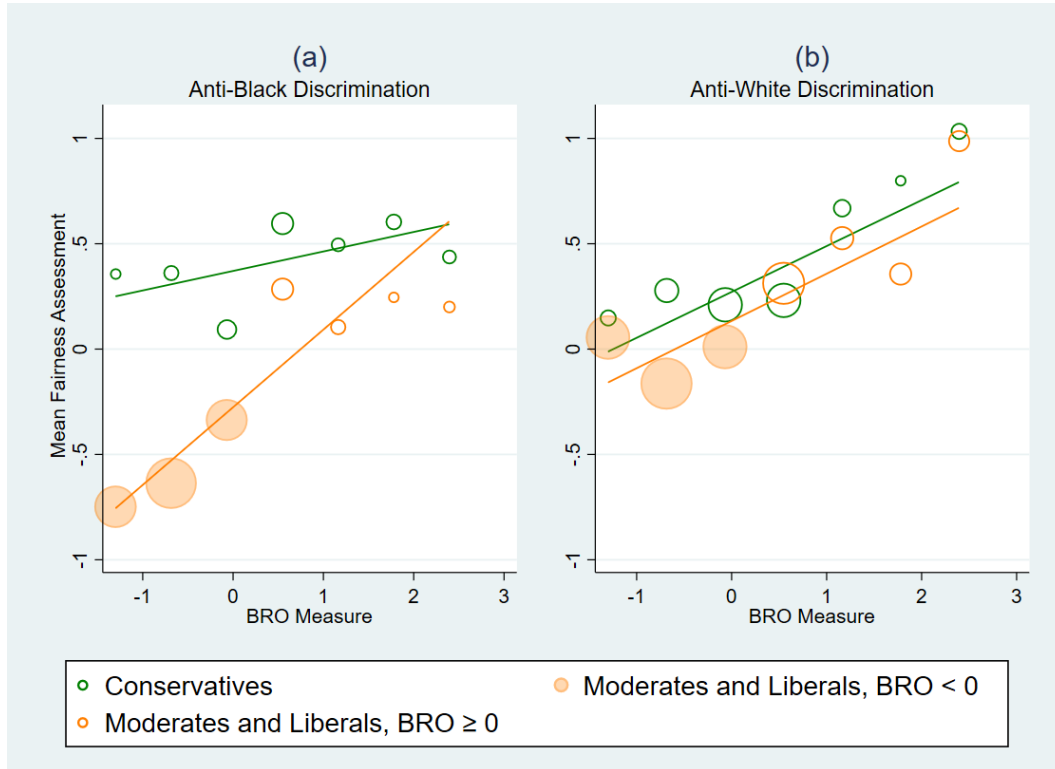
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields $p = .567$. A test for equality between conservatives and (moderates +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

Figure B11.6 Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 2.7)



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

Figure B11.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 2.8)



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

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

B.12 Replicating the Main Figures for ‘Thoughtful’ Subjects Only

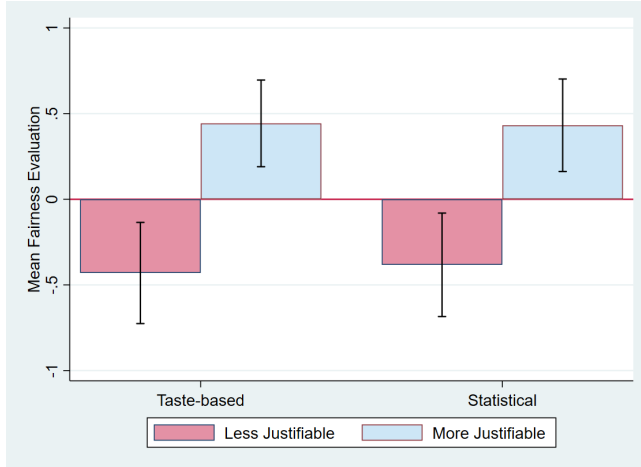
In this Appendix, we replicate Figures 2.2-2.8 with a subsample of “thoughtful” respondents. These respondents took more than the median amount of time (i.e., 8.37 minutes) to read our vignettes and think about their fairness assessments. This sample is composed of approximately 30% of respondents who identify as conservatives, 35% who identify as moderates, and 35% who identify as liberals. A comparison in the demographics between the full sample and subsample of thoughtful respondents is provided below in Table B12.1.

Table B12.1: Demographic Composition of MTurk Sample versus the American Community Survey (ACS)

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

Notes: This table is similar to Table B2.1, but it compares the full MTurk sample with the sub-sample of “thoughtful” respondents.

Figure B12.1: Fairness Ratings by Type of Discrimination and *Justifiability* (Replicates Figure 2.2)



Less vs More Justifiable Treatment

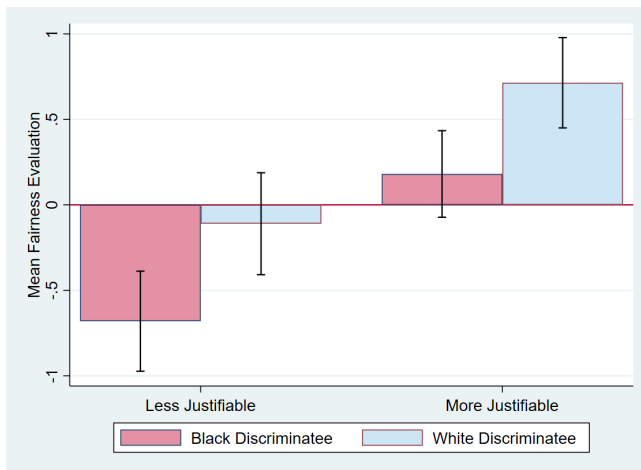
Overall: $p = .000$
 Within taste-based: $p = .000$
 Within statistical: $p = .000$

Taste vs Statistical Discrimination

Overall: $p = .918$
 Within Less-Justifiable: $p = .824$
 Within More-Justifiable: $p = .953$

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

Figure B12.2: Fairness by *Justifiability* and Discriminatee Race (replicates Figure 2.3)



Black vs White Treatment

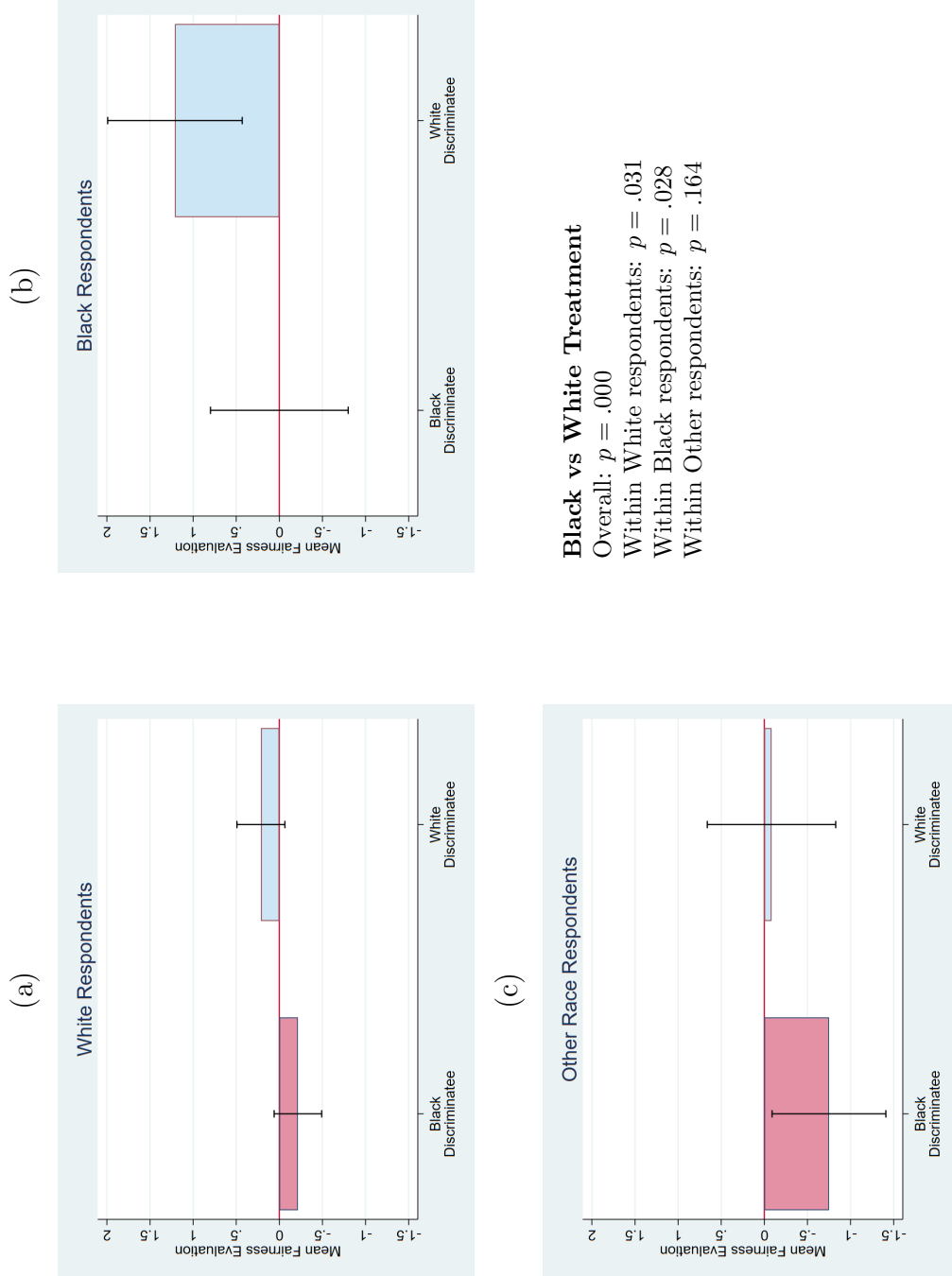
Overall: $p = .000$
 Within taste-based: $p = .007$
 Within statistical: $p = .004$

Less vs More Justifiable Treatment

Overall: $p = .000$
 Within Less-Justifiable: $p = .000$
 Within More-Justifiable: $p = .000$

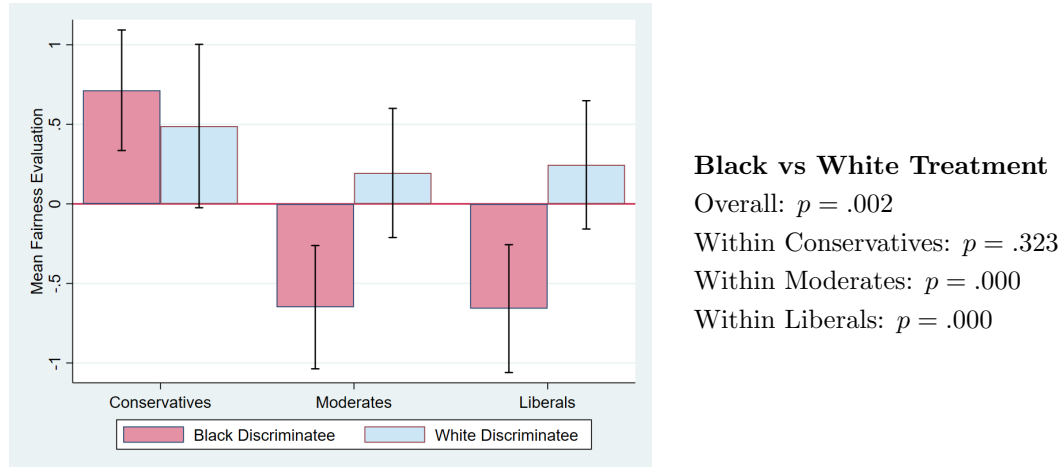
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. Within Black Discriminatees, less-justifiable scenarios are 0.861 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.825 units less fair. A test for equality of the Less versus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

Figure B12.3 Fairness Ratings by Respondent Race and Discriminatee Race (replicates Figure 2.4)



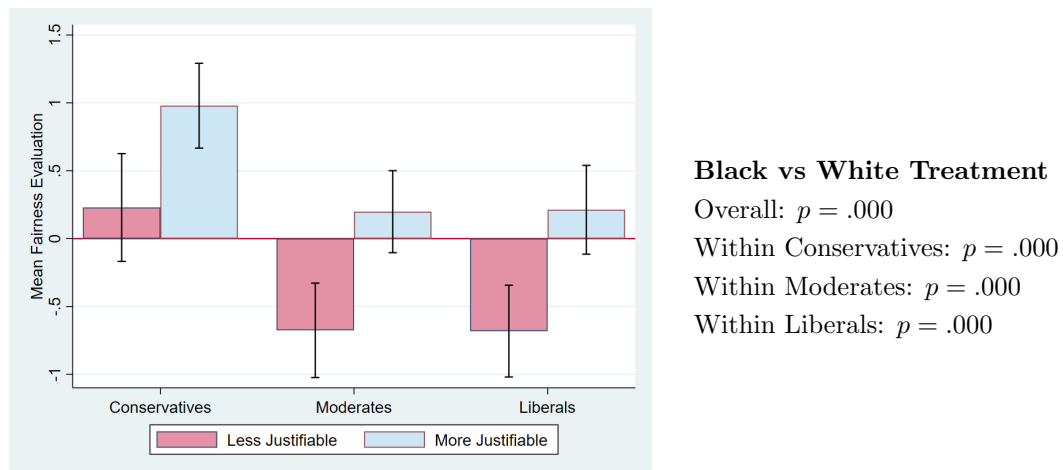
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. A test for equality of the discriminatee race effect (i.e. the Black treatment) across all three racial groups yields $p = .261$.

Figure B12.4: Fairness Ratings by Political Orientation and Discriminatee Race (replicates Figure 2.5)



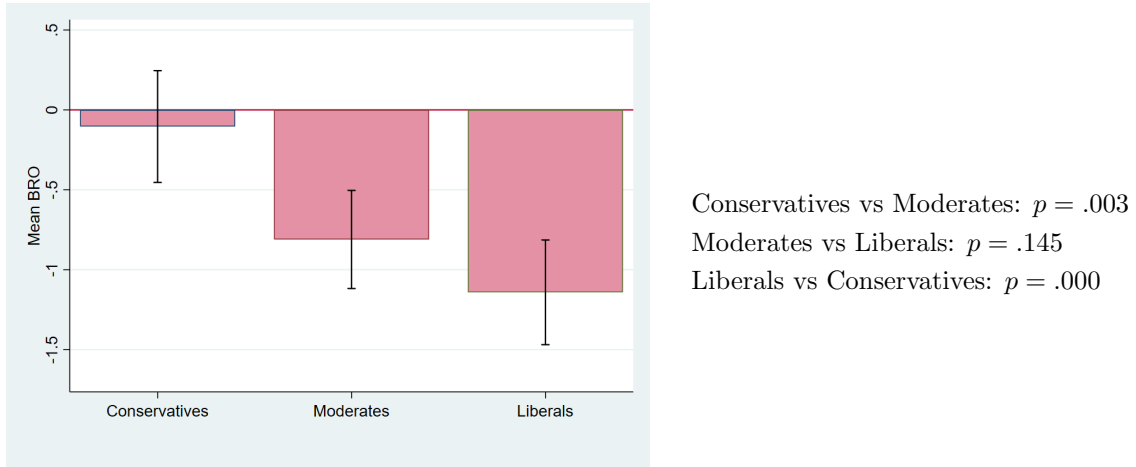
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields $p = .880$. A test for equality between conservatives and (moderates +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

Figure B12.5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent’s Political Leaning (replicates Figure 2.6)



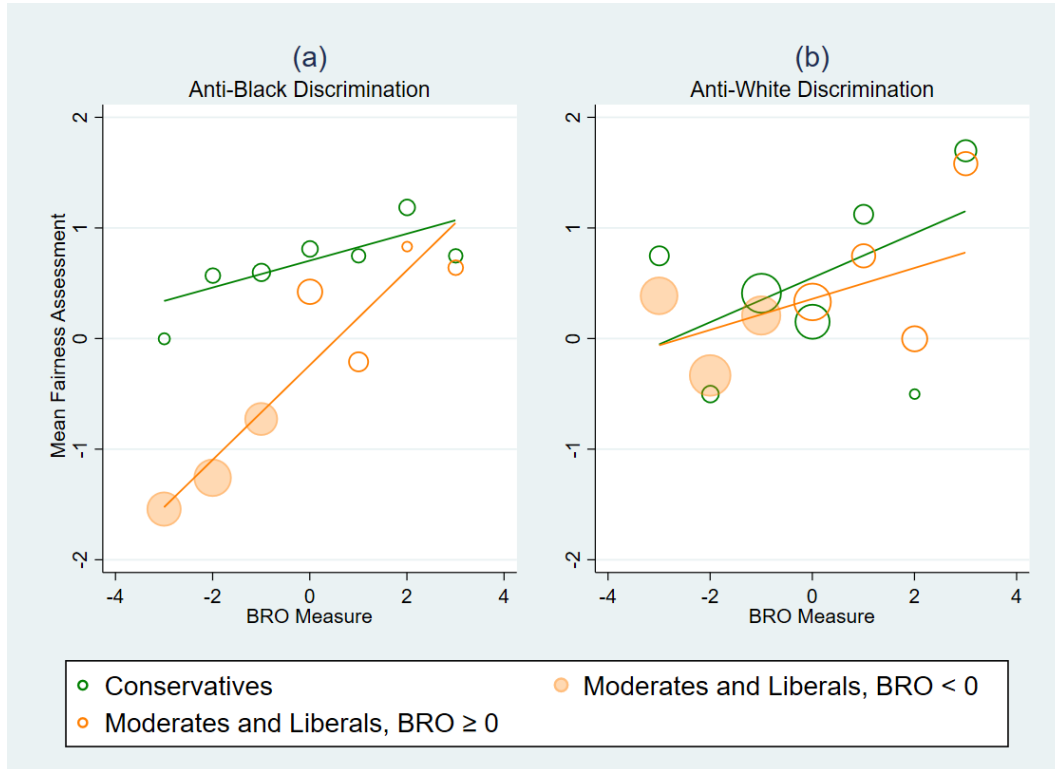
Notes: This figure is based on only Stage 1 observations. All p -values are clustered by respondent. A test for equality of the Less versus More Justifiability Gap across Conservatives, Moderates, and Liberals yields $p = .590$.

Figure B12.6 Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning (replicates Figure 2.7)



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

Figure B12.7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race (replicates Figure 2.8)



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

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

P Populated Pre-Analysis Plan

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

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

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

⁶Our 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>.

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

P1 Establish the Main Facts

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

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

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

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

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

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

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

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

In the paper, we use similar t-tests to compare the fairness of Statistical and Taste-Based Discrimination, as well as the sub-types of each (which we collectively call more-versus less-justifiable discriminatory acts) in Figure 2.2. The only difference from the PAP is that we restrict the sample to Stage 1 survey responses. This was to avoid possible contamination by the question order effects for the *race* treatment we discovered. The results are essentially identical to the PAP. We explored how the discriminatee race effect varies with the respondent’s own race in Figure 2.4 (which implements a similar t-test) and discuss the implications of our findings for the racial in-group bias model in Section 2.4.2. The in-group bias model is rejected in all cases.

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

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

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

$$BRO_i = \alpha + \theta^1 RR_i + \theta^2 RG_i + \theta^3 RA_i + \theta^4 RP_i + \epsilon_i \quad (\text{P.1})$$

where BRO_i is respondent i ’s assessment of Black peoples’ relative opportunities.⁹ RR , RG , RA , and RP represent (sets of) dummy variables for respondent race, gender, age, and political preferences, respectively. The PAP stated that we do not have strong priors for these effects, though we noted that factors like in-group bias could generate motivated

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

⁹Due to a cut-and-paste error, the PAP erroneously stated that equation P.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.

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 2.6 shows the relationship between the respondent's political leaning and BRO (essentially the θ^4 coefficients in equation (P.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.

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

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

Notes: This table contains the results of estimating equation (P.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:

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

where $FAIR_{ij}$ is respondent i 's assessment of the fairness of scenario j . In equation (P.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 2.4 displays the discriminatee race effect by respondent race (essentially, equation P.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 2.5 displays the discriminatee race effect by political orientation (essentially φ^4). As in Table P1.3, we find large differences: conservatives do not consider respondent race but moderates and liberals do.

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

	coefficient	s.e.
A. Treatment Effects:		
Taste-based	-0.0465	(0.0489)
Statistical \times Low-quality info	-0.490***	(0.0390)
Taste-based \times Customer	-0.474***	(0.0355)
Black discriminatee	0.0336	(0.130)
B. Respondent Characteristics		
Black respondent	0.0627	(0.125)
Other race respondent	-0.138	(0.122)
Male	0.0272	(0.0764)
Age 35-44	-0.0392	(0.0859)
Age 45-54	-0.0794	(0.108)
Age 55 and over	0.0461	(0.135)
Democrat	-0.233***	(0.0858)
Independent or other party	-0.320***	(0.121)
Liberal	-0.190*	(0.104)
Moderate	-0.131	(0.103)
Four-year college or university	0.240***	(0.0841)
Graduate school	0.433***	(0.107)
C. Race Treatment \times Respondent Characteristics		
Black Discriminatee \times Black respondent	0.0301	(0.102)
Black Discriminatee \times Other race respondent	0.0844	(0.148)
Black Discriminatee \times Male respondent	0.104	(0.0838)
Black Discriminatee \times Age 35-44	-0.0686	(0.102)
Black Discriminatee \times Age 45-54	0.0471	(0.111)
Black Discriminatee \times 55 and over	-0.300**	(0.140)
Black Discriminatee \times Democrat	-0.110	(0.0938)
Black Discriminatee \times Indep. or other party	0.0310	(0.1390)
Black Discriminatee \times Liberal	-0.270**	(0.114)
Black Discriminatee \times Moderate	-0.266**	(0.112)
Black Discriminatee \times Four-year college	0.0447	(0.0938)
Black Discriminatee \times Graduate school	-0.120	(0.118)
Constant	0.427**	(0.133)
Observations	2,568	
R-squared	0.169	

Notes: This table contains the results of estimating equation (P.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 (s.e.) 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:

$$\begin{aligned}
 FAIR_{ij} = & \alpha + \beta^1 T_{ij} + \beta^2 (S_{ij} \times L_{ij}) + \beta^3 (T_{ij} \times E_{ij}) \\
 & + \delta^1 RW_i + \delta^2 RB_i + \delta^3 (RW_i \times B_{ij}) + \delta^4 (RO_i \times B_{ij}) \\
 & + \delta^5 (RB_i \times B_{ij}) + \epsilon_{ij}
 \end{aligned} \tag{P.3}$$

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

As laid out in the PAP, Table P1.6 estimates equation (P.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:

¹⁰As 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.

(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.

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

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

Notes: This table contains the results of estimating equation (P.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 Preference 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:

$$FAIR_{ij} = \alpha + \beta A_{ij} + \delta B_{ij} + \epsilon_{ij} \quad (\text{P.4})$$

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 2.4.1. 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 2.4.3. 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 2.4.2. Our statistical power is too low to draw conclusions for non-White respondents, but (as in the PAP) we decisively reject it for White respondents.

Table P2.1: Assessing Three Models of Fairness

	All Respondents (1)	All Respondents (2)	White Respondents (3)	White Respondents (4)	Non-White Respondents (5)	Non-White Respondents (6)
Taste-based	-0.0498 (0.0492)	-0.0108 (0.0513)	-0.0833 (0.0559)	0.00662 (0.0592)	0.0669 (0.103)	-0.121 (0.197)
Statistical \times Low-quality	-0.490*** (0.0388)	-0.490*** (0.0449)	-0.531*** (0.0440)	-0.531*** (0.0508)	-0.344*** (0.0822)	-0.660*** (0.182)
Taste \times Employer	-0.474*** (0.0354)	-0.474*** (0.0408)	-0.462*** (0.0403)	-0.462*** (0.0465)	-0.514*** (0.0742)	-0.986*** (0.165)
Black Discriminatee	-0.181*** (0.0427)	-0.163*** (0.0378)	-0.191*** (0.0476)	-0.151*** (0.0430)	-0.145 (0.0964)	-0.374** (0.154)
Constant	0.358*** (0.0459)	0.0862** (0.0369)	0.379*** (0.0502)	0.0787* (0.0425)	0.283*** (0.108)	2.159*** (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 (P.4) from the pre-analysis plan. Columns 1-2 include all respondents, regardless of their race. Columns 3-4 only include White respondents. Finally, Columns 5-6 only include Non-white respondents. Two stars indicate a five percent significance level, and three stars indicate a one percent level. Standard errors are clustered by the respondent.

P2.4 A Hybrid Model: Conditional Utilitarianism

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

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

In equation (P.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 ($\delta^3 > 0$).

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

Table P2.4 contains our estimates of equation (P.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.

Table P2.4: Testing the Conditional Utilitarianism Model

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

Notes: This table contains the results of estimating equation (P.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 (P.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 (P.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 \times Low-quality, and Taste \times 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 2.5, 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 \times Low-quality	-0.941*** (0.0746)	-0.941*** (0.0861)	-0.970*** (0.0983)	-0.909*** (0.0965)
Taste-based \times 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	4.401*** (0.0881)	3.880*** (0.0707)	4.492*** (0.114)	4.306*** (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

Notes: Column 1 of Table P2.5 reproduces column 1 of Table P2.1. 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 2.7 so this is no longer relevant either. We also proposed working with more detailed racial identity categories, but (as expected) our samples were much too small for this.

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

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

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

Notes: This table compares estimates for equation (P.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 2.3.
- All four theoretical models of discrimination described in Section 2.4, 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 2.3 and 2.4 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 2.8's exploration of the "BRO hypothesis" we decided to use a continuous version of BRO (all seven values) rather than a dichotomized version, to show additional detail.

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

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

Appendix C

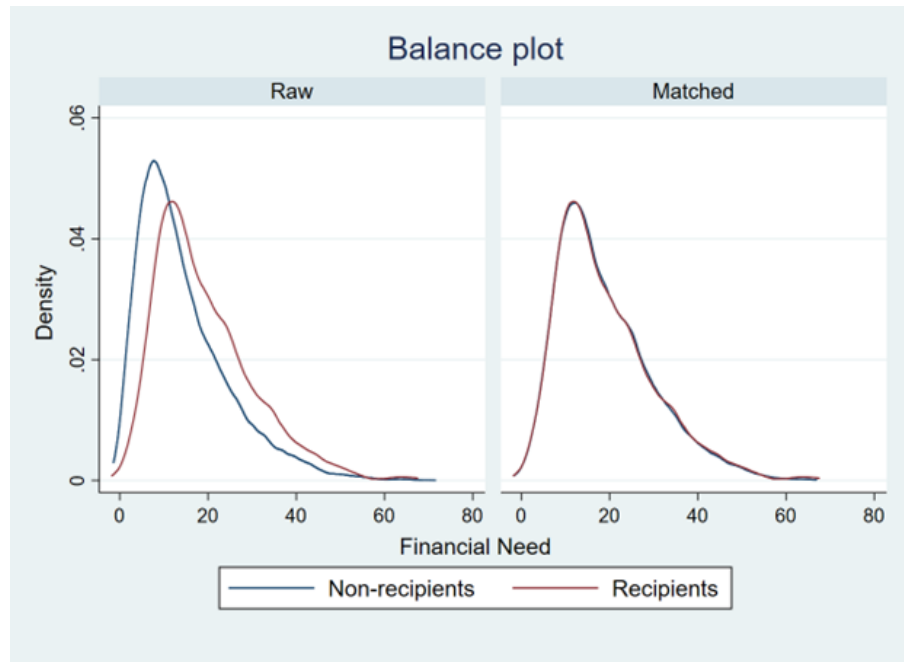
Additional Material for Chapter 3

C.1 More on B & B

Table C1.1 is analogous to Table 3.1, and it displays the summary statistics for students who were surveyed for the B & B follow-up studies. As discussed, students from the NPSAS surveys were randomly sampled to participate in the B & B. Consequently, the summary statistics among these individuals are mostly comparable to that of Table 3.1. The right-hand panel (Columns 1-3) reflects the full B & B sub-sample while the left (Columns 4-6) corresponds to the trimmed matching sub-sample, the latter of which results from non-matched students being discarded by the matching algorithm. The summary statistics between the full and trimmed sub-samples are very similar, which could be owed to B & B's representativeness.

Table C1.2 displays the summary statistics for covariates within the matched B & B sub-sample. It shows that these covariates are balanced across the recipient and non-recipient groups as a result of the NNM algorithm. All discrete covariates are perfectly balanced. The difference in financial need between recipients and non-recipients is also relatively small and insignificant ($p = .640$). Finally, as shown in Figure C1.1, the distributions for financial need for recipients and non-recipients are very similar and are indistinguishable ($p = 1.00$).

Figure C1.1: Distribution for Financial Need (B & B)



Notes: This figure is similar to Figure 3.1. However, it reflects the B & B sub-sample. It serves as a check for how well matching improves the balance of covariates, in particular, financial need. Here, the distributions between recipients and non-recipients in the matched sample are statistically indistinguishable ($p = 1.000$).

SOURCE: 2003-2004, 2007-2008, 2011-2012, & 2015-2016 NPSAS, and 2008-2009 & 2016-2017 B & B.

Table C1.1: Summary Statistics – Unmatched Sample (B & B sub-sample)

	Full Sample			Matching Sample		
	Recipients (1)	Non- recipients (2)	diff. <i>p</i> -val. (3)	Recipients (4)	Non- recipients (5)	diff. <i>p</i> -val. (6)
Worked at all	0.60 (0.49)	0.66 (0.47)	0.00	0.60 (0.49)	0.66 (0.47)	0.00
Hours worked while enrolled	12.04 (13.16)	14.34 (14.23)	0.00	11.93 (12.91)	13.91 (13.38)	0.00
Standardized GPA	0.24 (0.71)	0.36 (0.71)	0.00	0.26 (0.71)	0.37 (0.71)	0.00
STEM Major	0.52 (0.50)	0.47 (0.50)	0.00	0.53 (0.50)	0.48 (0.50)	0.01
Loan-forgiveness Occupation	0.15 (0.36)	0.16 (0.37)	0.29	0.15 (0.36)	0.17 (0.37)	0.12
Financial need	19.33 (11.05)	14.63 (10.46)	0.00	19.15 (10.88)	14.45 (10.28)	0.00
Underrepresented Minority	0.23 (0.42)	0.20 (0.40)	0.01	0.19 (0.39)	0.14 (0.45)	0.01
Female	0.59 (0.49)	0.58 (0.49)	0.70	0.59 (0.49)	0.60 (0.49)	0.57
Full-time status	0.74 (0.44)	0.69 (0.46)	0.00	0.77 (0.42)	0.73 (0.44)	0.00
Dependent	0.72 (0.45)	0.66 (0.47)	0.00	0.75 (0.42)	0.71 (0.44)	0.00
Seeking a Bachelor's degree	1.00 (0.00)	1.00 (0.03)	0.02	1.00 (0.00)	1.00 (0.00)	–
State Resident	0.76 (0.42)	0.84 (0.37)	0.00	0.79 (0.41)	0.87 (0.34)	0.00
Attends a private institution	0.55 (0.50)	0.39 (0.49)	0.00	0.54 (0.50)	0.39 (0.49)	0.00
Very selective institution	0.39 (0.49)	0.27 (0.44)	0.00	0.41 (0.49)	0.27 (0.44)	0.00
Moderately selective institution	0.53 (0.50)	0.62 (0.49)	0.00	0.56 (0.50)	0.69 (0.46)	0.00
Minimally selective institution	0.06 (0.24)	0.07 (0.26)	0.10	0.03 (0.18)	0.03 (0.18)	0.83
Open admissions institution	0.02 (0.14)	0.04 (0.19)	0.05	0.01 (0.08)	0.01 (0.09)	0.86
Observations	1,500	9,540	–	1,320	7,760	–

Notes: This table displays the summary statistics for all students included in the B & B sub-sample. It is arranged similarly to Table 1. Financial need is expressed in thousands and in 2016 USD. Sample sizes are rounded to the nearest tens per the IES's data security requests.

SOURCE: 2003-2004, 2007-2008, 2011-2012, & 2015-2016 NPSAS, and 2008-2009 & 2016-2017 B & B.

Table C1.2: Summary Statistics for Covariates in the Matched Sample (B & B)

	Recipients (1)	Non- recipients (2)	diff. <i>p</i> -val. (3)
Financial need	19.15 (10.88)	18.95 (10.53)	0.64
Underrepresented Minority	0.19 (0.39)	0.19 (0.39)	1.00
Female	0.59 (0.49)	0.59 (0.49)	1.00
Full-time status	0.77 (0.42)	0.77 (0.42)	1.00
Dependent	0.75 (0.43)	0.75 (0.43)	1.00
Seeking a Bachelor's degree	1.00 (0.06)	1.00 (0.06)	1.00
State Resident	0.79 (0.41)	0.79 (0.41)	1.00
Attends a private institution	0.54 (0.50)	0.54 (0.50)	1.00
Very selective institution	0.41 (0.49)	0.41 (0.49)	1.00
Moderately selective institution	0.56 (0.50)	0.56 (0.50)	1.00
Minimally selective institution	0.03 (0.18)	0.03 (0.18)	1.00
Open admissions institution	0.01 (0.08)	0.01 (0.08)	1.00
Observations	1,320	1,320	–

Notes: This table displays summary statistics for all covariates in the matched B & B sub-sample. Column 1 contains the means with standard deviations in parenthesis for Perkins loan recipients. Columns 2 contains those of non-recipients. Column 3 contains the *p*-values for the difference in means between these two groups. Financial need is expressed in thousands and 2016 USD. Sample sizes are rounded to the nearest tens per the IES's data security requests.

SOURCE: 2008-2009 and 2016-2017 B & B.

C.2 Predicting Perkins Take-up

Tables C2.1 and C2.2 display the results of estimating a linear probability model (LPM) of taking up Perkins loans on a set of covariates included in \mathbf{X}_i from equation (3.2). The latter table only incorporates students who were surveyed for the B & B. These tables show that the take-up of these loans is not random, i.e., it depends on various factors, such as financial need, dependency status, and institutional selectivity. The LPM is estimated across different sample restrictions for students' financial need, and these tables also show that these point estimates are mostly comparable across each one.

Table C2.1: Predicting Perkins Loan Take-up

	$need_i > 0$ (1)	$need_i > 5,000$ (2)	$need_i > 10,000$ (3)	$need_i > 15,000$ (4)
Financial need	0.00480*** (0.000297)	0.00367*** (0.000337)	0.00183*** (0.000407)	0.000477 (0.000521)
URM	0.00720 (0.00465)	0.00694 (0.00511)	0.00866 (0.00614)	0.00681 (0.00799)
Female	0.00145 (0.00328)	0.00143 (0.00367)	-0.00232 (0.00456)	-0.00100 (0.00603)
Full-time status	0.0191*** (0.00401)	0.0207*** (0.00454)	0.0188*** (0.00573)	0.0216*** (0.00770)
Dependent	0.0337*** (0.00453)	0.0411*** (0.00497)	0.0483*** (0.00598)	0.0591*** (0.00765)
Seeking a Bachelor's Degree	0.0404*** (0.00827)	0.0461*** (0.00948)	0.0645*** (0.0109)	0.0706*** (0.0147)
Resident	-0.0304*** (0.00644)	-0.0347*** (0.00689)	-0.0408*** (0.0749)	-0.0426*** (0.0866)
Attends a private institution	0.0333*** (0.00717)	0.0380*** (0.00779)	0.0524*** (0.00883)	0.0662*** (0.0103)
Institution is moderately selective	-0.0328*** (0.00831)	-0.0362*** (0.00919)	-0.0368*** (0.0105)	-0.0332*** (0.0126)
Institution is minimally selective	-0.0299*** (0.0113)	-0.0332*** (0.0127)	-0.0283* (0.0159)	-0.0147 (0.0213)
Institution has open admission	-0.0357*** (0.0108)	-0.0417*** (0.0121)	-0.0387*** (0.0145)	-0.0202 (0.0197)
Constant	0.0937*** (0.0167)	0.117*** (0.0190)	0.146*** (0.0220)	0.161*** (0.0280)
Survey Wave FE	YES	YES	YES	YES
Observations	54,080	47,470	34,350	21,820

Notes: This table displays the results of regressing an indicator for working a non-academic job on an indicator for taking up the Perkins loan. Each column corresponds to sample restrictions for financial need. One star indicates a 10% significance level, two stars indicate a 5% level, and three stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution-by-year cells. Sample sizes are rounded to the nearest tens per the IES's data security requests.

SOURCE: 2003-2004, 2007-2008, 2011-2012, & 2015-2016 NPSAS, and 2008-2009 & 2016-2017 B & B.

Table C2.2: Predicting Perkins Loan Take-up (B& B Sample)

	$need_i > 0$ (1)	$need_i > 5,000$ (2)	$need_i > 10,000$ (3)	$need_i > 15,000$ (4)
Financial need	0.00430*** (0.000497)	0.00326*** (0.000574)	0.00145** (0.000724)	0.00128 (0.000919)
URM	0.0147 (0.00986)	0.0164 (0.0109)	0.0250* (0.0136)	0.0342** (0.0171)
Female	0.00227 (0.00761)	0.00486 (0.00859)	-0.00186 (0.0104)	-0.0150 (0.0128)
Full-time status	0.00312 (0.00836)	0.00356 (0.00954)	-0.00300 (0.0126)	-0.000887 (0.0166)
Dependent	0.00119 (0.00815)	0.00594 (0.00921)	0.0135 (0.0117)	0.0122 (0.0153)
Seeking a Bachelor's degree	0.0842*** (0.0163)	0.0103*** (0.0160)	0.01306*** (0.0275)	0.0133*** (0.0276)
Resident	-0.0229** (0.0113)	-0.0261** (0.0122)	-0.0360*** (0.0139)	-0.0426*** (0.0160)
Attends a private institution	0.0217* (0.0124)	0.0248* (0.0136)	0.0364** (0.0161)	0.0428** (0.0185)
Institution is moderately selective	-0.0578*** (0.00139)	-0.0634*** (0.00155)	-0.0703*** (0.0184)	-0.0572*** (0.0210)
Institution is minimally selective	-0.0582*** (0.0192)	-0.0634*** (0.0220)	-0.0789*** (0.0269)	-0.0799** (0.0322)
Institution has open admission	-0.0665*** (0.0200)	-0.0755*** (0.0223)	-0.979*** (0.0258)	-0.0586 (0.0416)
Constant	0.0419 (0.0272)	0.00487* (0.0285)	0.0771* (0.0395)	0.0795* (0.0416)
Survey Wave FE	YES	YES	YES	YES
Observations	11,040	9,600	6,870	4,500

Notes: This table is similar to Table C2.1, but it reflects the B & B sub-sample. Each column corresponds to sample restrictions for financial need. One star indicates a 10% significance level, two stars indicate a 5% level, and three stars indicate a 1% level. Standard errors are in parenthesis and clustered by institution-by-year cells. Sample sizes are rounded to the nearest tens per the IES's data security requests.

SOURCE: 2003-2004, 2007-2008, 2011-2012, & 2015-2016 NPSAS, and 2008-2009 & 2016-2017 B & B.

Bibliography

- Abadie, Alberto and Guido Imbens**, “Large Sample Properties of Matching Estimators For Average Treatment Effects,” *Econometrica*, 2006, 74 (1), 235–267.
- and –, “Bias-Corrected Matching Estimators for Average Treatment Effects,” *Journal of Business & Economic Statistics*, 2011, 29 (1), 1–11.
- Abeler, Johannes, Sebastian Kube, Steffen Altmann, and Matthias Wibrals**, “Gift Exchange and Workers’ Fairness Concerns: When Equality is Unfair,” *Journal of the European Economic Association*, 2010, 8 (6), 1299 – 1324.
- Alesina, Alberto and Eliana La Ferrara**, “Preferences for Redistribution in the Land of Opportunities,” *Journal of Public Economics*, 2005, 89, 897 – 931.
- , **Armando Miano, and Stefani Stantcheva**, “The Polarization of Reality,” *American Economic Review Papers and Proceedings*, 2020, 110, 324 – 328.
- , **Matteo F. Ferroni, and Stefani Stantcheva**, “The Polarization of Reality,” Working Paper 29245, National Bureau of Economic Research 2021.
- Almås, Ingvild, Alexander Cappelen, and Bertil Tungodden**, “Cutthroat Capitalism versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking than Scandinavians?,” *Journal of Political Economy*, 2020, 128 (5), 1753 – 1788.
- Alon, S.**, “The evolution of class inequality in higher education: Competition, exclusion, and adaption,” *American Sociological Review*, 2009, 74 (5), 731–755.
- Andreoni, James, Deniz Aydin, Blake Barton, B. Douglas Bernheim, and Jeffrey Naecker**, “When Fair Isn’t Fair: Understanding Choice Reversals Involving Social Preferences,” *Journal of Political Economy*, 2020, 128 (5), 1673 – 1711.
- Arcidiacono, Peter, Estaban M. Aucejo, and Ken Spenner**, “What happens after enrollment? An analysis of the time path of racial differences in GPA and major choice,” *IZA Journal of Labor Economics*, 2012, 1 (5), 1–24.

- , **Esteban M. Aucejo, and Joseph V. Hotz**, “University Differences in the Graduation of Minorities in STEM Fields: Evidence from California,” *American Economic Review*, 2016, *106* (3), 525 – 562.
- , **Michael Lovenheim, and Maria Zhu**, “Affirmative Action in Undergraduate Education,” *Annual Review of Economics*, 2015, *7*, 487 – 518.
- Arechar, Antonio A., Simon Gächter, and Lucas Molleman**, “Conducting Interactive Experiments Online,” *Experimental Economics*, 2018, *21*, 99 – 131.
- Astin, Alexander W.**, “How Liberal Arts Colleges Affect Students,” *Daedalus*, 1999, *128* (1), 77 – 100.
- Auspurg, Katrin, Thomas Hinz, and Karsten Sauer**, “Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments,” *American Sociological Review*, 2017, *82* (1), 179 – 210.
- Avery, Christopher and Jonathan Levin**, “Early Admissions at Selective Colleges,” *American Economic Review*, 2010, *100* (5), 2125 – 2156.
- Backes, Ben**, “Do Affirmative Action Bans Lower Minority College Enrollment and Attainment? Evidence from Statewide Bans,” *Journal of Human Resources*, 2012, *47* (2), 435 – 455.
- Barr, Abigail, Tom Lane, and Daniele Nosenzo**, “On the Social Inappropriateness of Discrimination,” *Journal of Public Economics*, 2018, *164*, 153 – 164.
- Becker, Gary S.**, *The Economics of Discrimination (second edition)*, Chicago: University of Chicago Press, 1971.
- Belasco, Andrew S., Kelly O. Rosinger, and James C. Hearn**, “The Test Optional Movement at America’s Selective Liberal Arts Colleges: A Boon for Equity or Something Else?,” *Education Evaluation and Policy Analysis*, 2015, *37* (2), 206 – 223.
- Bennett, C.T.**, “Untested admissions: Examining changes in application behaviors and student demographics under test-optional policies,” *American Educational Research Journal*, 2022, *59*, 180–216.
- Bertrand, Marianne and Esther Duflo**, “Field Experiments on Discrimination,” in Abhijit Vinayak Banerjee and Esther Duflo, eds., *Handbook of Economic Field Experiments*, Elsevier, 2017, pp. 309 – 383.
- Bettinger, Eric**, “How Financial Aid Affects Persistence,” In *College choices: The economics of where to go, when to go, and how to pay for it*, 2004, pp. 207–238.
- Blau, J.R., S. Moller, and LV Jones**, “Why test? Talent loss and enrollment loss,” *Social Science Research*, 2004, *33* (3), 409 – 434.

- Bohren, J. Aislin, Kareem Haggag, Alex Imas, and Devin G. Pope**, “Inaccurate Statistical Discrimination,” Working Paper 25447, National Bureau of Economic Research 2019.
- Bound, John, Brad Hershbein, and Bridget Terry Long**, “Playing the admissions game: Student reactions to increasing college competition,” *Journal of Economics Perspectives*, 2009, 23 (4), 119 – 146.
- , **Michael F. Loveheim, and Sarah E. Turner**, “Why Have College Completion Rates Declines? An Analysis of Changing Student Preparation and Collegiate Resources,” *American Economic Journal: Applied Economics*, 2010, 2 (3), 129–157.
- Bracha, Anat, Uri Gneezy, and George Loewenstein**, “Relative Pay and Labor Supply,” *Journal of Labor Economics*, 2015, 33 (2), 297 – 315.
- Breza, Emily, Supreet Kaur, and Yogita Shamdasani**, “The Morale Effects of Pay Inequality,” *Quarterly Journal of Economics*, 2017, 133 (2), 611 – 663.
- Broton, Katharine M., Sara Goldrick-Rab, and James Benson**, “Working for college: The causal impacts of financial grants on undergraduate employment,” *Education Evaluation and Policy Analysis*, 2016, 38 (3), 477–494.
- Bruhin, Adrian, Ernst Fehr, and Daniel Schunk**, “The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences,” *Journal of the European Economic Association*, 2019, 17 (4), 1025 — 1069.
- Cain, Glen G.**, “The Economic Analysis of Labor Market Discrimination: A Survey,” in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, Elsevier, 1986, pp. 693 – 781.
- Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez**, “Inequality at work: The effect of peer salaries on job satisfaction,” *Journal of Economic Behavior & Organization*, 2012, 102 (6), 2981 – 3003.
- **and Alan B. Krueger**, “Would the elimination of affirmative action affect highly qualified minority applicants? Evidence from California and Texas,” *ILR Review*, 2005, 58 (3), 416–434.
- Carnevale, Anthony P., Ban Cheah, and Martin Van Der Werf**, “ROI of Liberal Arts Colleges: Value Adds Up Over Time,” *Georgetown University Center on Education and the Workforce*, 2020, (14).
- Charness, Gary and David I. Levine**, “When Are Layoffs Acceptable? Evidence from a Quasi- Experiment,” *Industrial and Labor Relations Review*, 2000, 53 (3), 381 – 400.

- **and Peter Kuhn**, “Does Pay Inequality Affect Worker Effort? Experimental Evidence,” *Journal of Labor Economics*, 2007, 25 (4), 693 – 724.
- , **Till Gross**, and **Christopher Guo**, “Merit Pay and Wage Compression with Productivity Differences and Uncertainty,” *American Economic Review*, 2015, 117, 233 – 247.
- Chen, Y. and S. X. Li**, “Group identity and social preferences,” *American Economic Review*, 2009, 99 (1), 431 – 457.
- Cohn, Alain, Ernst Fehr, and Lorenze Götte**, “Fair Wages and Effort Provisions: Combining Evidence from a Choice Experiment and a Field Experiment,” *Management Science*, 2014, 61 (8), 1777 – 1794.
- Cullen, Zoe B. and Bobak Pakzad-Hurson**, “Equilibrium Effects of Pay Transparency,” 2017. Unpublished.
- Davidai, S. and J. Walker**, “Americans Misperceive Racial Disparities in Economic Mobility,” *Personality and Social Psychology Bulletin*, 2021, 48 (5), 793 – 806.
- de Chaisemartin, Clément and Xavier D’Haultfoeuille**, “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects,” *American Economics Review*, 2020, 110 (9), 2964—2996.
- **and** – , “Difference-in-Differences Estimators of Intertemporal Treatment Effects,” 2021. Working paper.
- Denning, Jeffrey T.**, “Born under a Lucky Star: Financial Aid, College Completion, Labor Supply, and Credit Constraints,” *Journal of Human Resources*, 2019, 54 (3), 760–784.
- , **Benjamin M. Marx**, and **Lesley J. Turner**, “ProPelled: The Effects of Grants on Graduation, Earnings, and Welfare,” *American Economic Journal: Applied Economics*, 2019, 11 (3), 193–224.
- Dynarski, Susan**, “Does aid matter? Measuring the effect of student aid on college attendance and completion,” *American Economic Review*, 2003, 93 (1), 279 – 288.
- , “Loans, Liquidity and Schooling Decisions,” Working Paper, Kennedy School of Government 2005.
- Emrey-Arras, Melissa**, “Financial Aid Offers: Action Needed to Improve Information on College Costs and Student Aid,” Government Report GAO-23-104708, U.S. Government Accountability Office, Washington, D.C. 2023.
- Epstein, Jonathan**, “Behind the SAT-Optional Movement: Context and Controversy,” *Journal of College Admission*, 2009, 204, 8–19.

- Everett, Jim A.C., Nadira S. Faber, and Molly Crockett**, “Preferences and Beliefs in Ingroup Favoritism,” *Frontiers in Behavioral Neuroscience*, 2015, 9, 15.
- Feess, E, J. Feld, and S. Noy**, “People Judge Discrimination Against Women More Harshly Than Discrimination Against Men – Does Statistical Fairness Discrimination Explain Why?,” *Frontiers in Psychology*, 2021, 12, 675776.
- Fehr, Dietman, Hannes Rau, Stefan T. Trautmann, and Yilong Xu**, “Fairness Properties of Compensation Schemes,” 2021. Unpublished.
- Fong, C. M. and E. F. Luttmer**, “What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty,” *American Economic Journal: Applied Economics*, 2009, 1 (2), 64 – 87.
- **and –**, “Do fairness and race matter in generosity? Evidence from a nationally representative charity experiment,” *Journal of Public Economics*, 2011, 95 (5), 372 – 394.
- Frank, Robert H.**, “Are Workers Paid Their Marginal Product?,” *American Economic Review*, 1984, 74 (4), 549 – 571.
- Freedle, R.O.**, “Correcting the SAT’s ethnic and social-class bias: A method for re-estimating SAT Scores,” *Harvard Education Review*, 2003, 73, 1 – 43.
- Gartenberg, Claudine and Julie Wulf**, “Pay Harmony? Social Comparison and Performance Compensation in Multibusiness Firms,” *Organization Science*, 2017, 28 (1), 39 – 55.
- Gervais, Martin and Nicolas L. Ziebarth**, “Life After Debt: Postgraduate Consequences of Federal Student Loans,” *Economic Inquiry*, 2019, 57 (3), 1342–1366.
- Goodman-Bacon, Andrew**, “Difference-in-Differences with Variation in Treatment Timing,” *Journal of Econometrics*, 2021, 225, 254—277.
- Haaland, C. and C. Roth**, “Beliefs About Racial Discrimination and Support For Pro-Black Policies,” *Review of Economics and Statistics*, 2021, pp. 1 – 15.
- Heckman, James T., Hidehiko Ichimura, and Petra Todd**, “Matching as an Econometric Evaluation Estimator,” *Review of Economic Studies*, 1997, 65 (2), 261 – 294.
- Hedegaard, Morten Størling and Jean-Robert Tyran**, “The Price of Prejudice,” *American Economic Journal: Applied Economics*, 2018, 10 (1), 40 – 63.
- Heller, D.E.**, “The Impact of Loans on Student Access,” in Donald E. Heller, Sandy Baum, Michael McPherson, and Patricia Steele, eds., *The Effectiveness of Student Aid Policies: What the Research Tells Us*, New York: College Board, 2008, pp. 39–68.

- Hinrichs, Peter**, “The Effects of Affirmative Action Bans on College Enrollment, Education Attainment, and the Demographic Compositions of Universities,” *Review of Economics and Statistics*, 2012, *94* (3), 712–722.
- Imbens, Guido**, “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *The Review of Economics and Statistics*, 2004, *86* (1), 4 – 29.
- and **Jeffrey Wooldridge**, “Recent developments in the econometrics of program evaluation,” *Journal of Economic Literature*, 2009, *47* (1), 5–86.
- Jacobsen, Louis, Robert Lalonde, and Danielle Sullivan**, “Earnings Losses of Displaced Workers,” *American Economic Review*, 1993, *83* (4), 685–709.
- Jasso, Guillermina and Peter H. Rossi**, “Distributive Justice and Earned Income,” *American Sociological Review*, 1977, *42* (4), 639 – 665.
- , **Robert Shelly, and Murry Webster**, “How Impartial are the Observers of Justice Theory?,” *Social Science Research*, 2019, *79*, 226 – 246.
- Kalenkoski, Charlene Marie and Sabrina Wulff Pabilonia**, “Parental transfers, student achievement, and the labor supply of college students,” *Journal of Population Economics*, 2022, *23*, 469 – 496.
- Kofoed, Michael**, “Pell Grants and Labor Supply: Evidence from a Regression Kink,” *IZA Discussion Paper No. 15061*, 2022.
- Kraus, M.W., I.N. Onyeador, N.M. Daumeyer, J.M. Rucker, and J.A. Richeson**, “The Misperception of Racial Economic Inequality,” *Perspectives on Psychological Sciences*, 2019, *14* (6), 899 – 921.
- , **J.M. Rucker, and J.A. Richeson**, “Americans misperceive racial economic equality.” *Proceedings of the National Academy of Sciences*, 2019, *114* (39), 10324 – 10331.
- Krupka, Erin L. and Roberto A. Weber**, “Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?,” *Journal of the European Economic Association*, 2013, *11* (3), 495 – 524.
- Kuhn, Peter and Trevor Osaki**, “When Is Discrimination Unfair?” 2020. AEA RCT Registry: September 22.
- Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva**, “How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments,” *American Economic Review*, 2015, *105* (4), 1478 – 1508.
- Lefgren, Lars J., David Sims, and Olga Stoddard**, “Effort, luck, and voting for redistribution,” *Journal of Public Economics*, 2016, *143*, 89 – 97.

- Lippens, Louis, Stijn Baert, and Eva Derous**, “Loss Aversion in Taste-Based Employee Discrimination: Evidence from a Choice Experiment,” 2021. *IZA Discussion Paper no. 14438*.
- Marin, Patricia and Catherine L. Horn**, “Realizing Bakke’s Legacy: Affirmative Action, Equal Opportunity, and Access to Higher Education,” 2008.
- Mas, Alexandre**, “Does Transparency Lead to Pay Compression,” *Journal of Political Economy*, 2017, 125 (5), 1638 – 1721.
- McPherson, Michael S. and Morton Owen Shapiro**, “The Future Economic Challenges for the Liberal Arts Colleges,” *Daedalus*, 1999, 128 (1), 47 – 75.
- Oprea, Ryan and Sevgi Yuksel**, “Does Transparency Lead to Pay Compression,” *Journal of the European Economic Association*, 2022, 20 (2), 667 – 699.
- Page, Lindsay C. and Judith Scott-Clayton**, “Improving college access in the United States: Barriers and policy responses,” *Economics of Education Review*, 2015, 51, 493–520.
- Pallais, Amanda**, “Small Differences That Matter: Mistakes in Applying to College,” *Journal of Labor Economics*, 2016, 33 (2), 4–22.
- Park, Rina Seung Eun and Judith Scott-Clayton**, “The Impact of Pell Grant Eligibility on Community College Students’ Financial Aid Packages, Labor Supply, and Academic Outcomes,” *Education Evaluation and Policy Analysis*, 2018, 40 (4), 557 – 585.
- Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer**, “Data quality of platforms and panels for online behavioral research,” *Behavioral Research Methods*, 2021, 54, 1643 – 1662.
- Pop-Echeles, Cristian and Miguel Urquiola**, “Going to a Better School: Effects and Behavioral Responses,” *American Economic Review*, 2013, 103 (4), 1289 – 1324.
- Rothstein, Jesse**, “College performance predictions and the SAT,” *Journal of Econometrics*, 2004, 121 (1-2), 297–317.
- **and Albert H Yoon**, “Affirmative Action in Law School Admissions: What Do Racial Preferences Do?,” *The University of Chicago Law Review*, 2008, 75, 649–714.
- Saboe, M. and S. Terrizzi**, “SAT optional policies: Do they influence graduate quality, selectivity or diversity?,” *Economic Letters*, 2019, 174, 13–17.
- Santelices, M.V. and Mark Wilson**, “Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning,” *Harvard Educational Review*, 2010, 80 (1), 106 – 134.

- Schildberg-Hörisch, Hannah, Marco A. Schwarz, Chi Trieu, and Jana Willrodt**, “Perceived Fairness and Consequences of Affirmative Action Policies,” 2022. CESifo Working Paper no. 10198.
- Scott-Clayton, Judith**, “What Explains Trends in Labor Supply Among U.S. Undergraduates?,” *National Tax Journal*, 2012, *65* (1), 181–210.
- Seftor, Neil S. and Sarah E. Turner**, “Back to school: Federal student aid policy and adult college enrollment,” *Journal of Human Resources*, 2002, *37* (2), 336 – 352.
- Spence, Michael**, “Job Market Signaling,” *Quarterly Journal of Economics*, 1974, *87* (3), 281 – 306.
- Stantcheva, Stefanie**, “Understanding Tax Policy: How do People Reason?,” *Quarterly Journal of Economics*, 2021, *136* (4), 2309 – 2369.
- Stinebrickner, Ralph and Todd R. Stinebrickner**, “Working During School and Academic Performance,” *Journal of Labor Economics*, 2003, *21* (1), 473–491.
- and –, “Time-use and College Outcomes,” *Journal of Econometrics*, 2004, *121* (1), 243–269.
- Sun, Liyang and Sarah Abraham**, “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*, 2021, *225*, 175–199.
- Tilcsika, András**, “Statistical Discrimination and the Rationalization of Stereotypes,” *American Sociological Review*, 2021, *86* (1), 93 – 122.
- Turner, Lesley J.**, “The Economic Incidence of Federal Student Grant Aid,” Working Paper 2017.
- Turner, Lesley J. and Benjamin M. Marx**, “Borrowing Trouble? Human Capital Investment with Opt-In Costs and Implications for the Effectiveness of Grant Aid,” *American Economic Journal: Applied Economics*, 2018, *10* (2), 163 – 201.