Racial discrimination in transportation network companies

Racial Discrimination in Transportation Network Companies

Yanbo Ge, Christopher R. Knittel,
Don MacKenzie, and Stephen Zoepf *

June 27, 2020

Abstract

In a randomized audit study, we sent passengers in Boston, MA on nearly 1,000 rides on controlled routes using the Uber and Lyft smartphone apps, recording key performance metrics. Passengers randomly selected between accounts that used African American-sounding and white-sounding names. We find that the probability an Uber driver accepts a ride, sees the name, and then cancels doubles when passengers used the account attached to the African American-sounding name. In contrast, Lyft drivers observe the name before accepting a ride and, as expected, we find no effect of name on cancellations. We do not, however, find that the increase in cancellations leads to measurably longer wait times for Uber.

Keywords: discrimination; audit study; field experiment; transportation network company; peer economy; sharing economy.
JEL Codes: J15, J16, L90, R40.

*Authors are listed alphabetically. This paper has benefited from seminar participants at Brown University, MIT, Stanford University, University of California, Berkeley, and University of Washington. We also thank 16 anonymous research assistants that served as confederates in the experiment, and Jim Anderson for tireless administrative support. This work was supported by the University of Washington Royalty Research Fund [grant number A98088]; and the MIT Center for Energy and Environmental Policy Research. *Ge:* Post-doctoral Researcher, Data Science in Energy Systems, National Renewable Energy Laboratory (Beginning May 2019), yanboge@uw.edu. *Knittel:* George P. Shultz Professor, Sloan School of Management, and Director of the Center for Energy and Environmental Policy Research, MIT, and NBER, knittel@mit.edu. *MacKenzie:* Associate Professor, Department of Civil and Environmental Engineering, University of Washington, dwhm@uw.edu, Box 352700, Seattle WA 98195-2700. *Zoepf:* Lecturer, Department of Mechanical Engineering, Stanford University, szoepf@stanford.edu.

1 Introduction

As technological and market advances of the peer economy continue to outpace regulation, enforcement of responsible behavior within peer economies has been left to internal feedback mechanisms and, occasionally, to the courts. One area of behavior with important social equity implications is that of discrimination in the provision of services. In the transportation sector, Transportation Network Companies (TNCs, also referred to as "ridesourcing" or "ridehailing" companies), such as Uber and Lyft, match individual travelers with drivers in real time.

Discrimination by taxi drivers has long been an acknowledged social issue, providing plenty of pop-culture fodder. Taxi drivers in most cities are required to pick up any passenger while on duty, and taxi drivers are reminded of this obligation (Harshbarger, 2015). Despite periodic high-profile incidents of taxi discrimination (e.g., Donnelly, 2015; Gonen, 2015; Glanville, 2015), equal-service provisions are difficult to enforce.

A growing body of research finds that TNC services have increased welfare for travelers (Cohen et al., 2016) and drivers (Hall and Krueger, 2018; Chen et al., 2017; Angrist et al., 2017), while using vehicles and drivers' time more efficiently (Cramer and Krueger, 2016). However, this paper is concerned with the equity of service quality by TNC companies, and specifically whether there is evidence that TNC drivers discriminate against African American passengers.

The relationship between TNCs and discrimination is a complex one, and our work complements other recent work in this area. At a high level, a study funded by Uber (Smart et al., 2015) found that UberX provided lower fares and shorter waits than taxis in areas of Los Angeles with below-average incomes. Hughes and MacKenzie (2016) found that expected waiting times for an UberX ride were shorter in Seattle-area neighborhoods that had lower average incomes and more minorities, even after adjusting for differences in residential and employment density. Our work complements Smart et al. (2015) by focusing on differences in TNC services between travelers of different races, instead of differences between taxis and UberX. While Hughes and MacKenzie (2016) focus on differences in expected waiting times across different neighborhoods, our work focuses on actual individual-level waiting times and cancellations. Our work also complements Brown (2018), which compared phone-dispatched taxis with Uber and Lyft in Los Angeles. While Uber's advertising certainly includes minority customers (e.g., Figure 1), popular press articles (e.g. Nicholson, 2013) have noted that TNC services are unavailable to customers without access to a credit card, who are more likely to be lower-income and members of minorities.

While these systemic concerns may be valid, TNCs have a limited ability to control them, as decisions about where to drive and whom to pick up are ultimately made by individual drivers. Drivers are required to maintain adequate overall measures of performance including a high rating from passengers, high ride acceptance rate, and low cancellation rate. However, even these driver-specific performance measures may not detect whether a driver behaves differently toward passengers based on their race or gender.

In this paper, we report on a field experiment undertaken in Boston, Massachusetts, USA, designed to test for discrimination in the ridehailing market. The primary question we seek to answer is whether the rapidly-growing TNC market treats customers of different races equally. To answer this, we hired research assistants (RAs) to serve as travelers summoning rides. Each RA requested rides under two different names: using the nomenclature of Fryer and Levitt (2004), one name was "white-sounding" while the other was a "distinctively black name."[1]

For each trip, the RAs gathered data using four time-stamped screenshots on their smartphones: (i) just before requesting a trip (with expected wait time); (ii) just after a trip is accepted by a driver (with a new wait time); (iii) when a driver arrives for pickup; and (iv) when the vehicle stops at the end of the trip. Using the data gathered from these screenshots, we evaluate waiting times, travel times, driver cancellations, costs, and (where applicable) ratings awarded by drivers to the travelers.

We find significant evidence of racial discrimination. We find that the probability a driver accepts a ride, but then subsequently cancels the ride, more than doubles for African American riders of UberX. One of the key differences between the Uber and Lyft apps, from the driver's perspective, is that the Lyft app shows the names and photos of the of the riders prior to trip acceptance, whereas Uber drivers only see the passenger's name after accepting a request. As expected, we find no effect on cancellations for African American riders of Lyft because, we surmise, given that names and photos are visible to the driver prior to acceptance, any discrimination occurs prior to accepting the initial request. We also show that this increase in the rate of cancellations is concentrated among African American males; their cancellation rates are three times that of white males. Furthermore, this cancellation effect is concentrated in areas with low population density, perhaps because drivers in those areas self-select to reduce their interaction with African Americans. We do not, however, find that the increase in cancellation rates manifests itself into increases in wait times.

The design of the Boston experiment was informed by an earlier pilot study that we conducted in Seattle, Washington, USA. In the pilot, half of the RAs were African American and half were white, and requested trips on Uber, Lyft, and Flywheel (a taxi-hailing app) using their actual names and profile photos. We did not systematically record cancellations during the Seattle pilot. The results of that pilot were suggestive of longer waiting times for African Americans, though the differences in waiting times were not statistically significant. The longer wait times in Seattle, which were not replicated in the Boston experiment, may reflect the thicker market that exists in Boston, with cancellations being more quickly rerouted to a nearby driver. However, the difference in results between the two cities may be an artifact of the Seattle pilot being underpowered and using a between-subjects design, in contrast to the within-subjects design employed in Boston. Based on our experience in the Seattle pilot, we made two major changes before conducting the Boston experiment.

[1] Bertrand and Mullainathan (2004) found that when stereotypically white names and stereotypically African American names were randomly assigned to job applications, those with the white names received 50% more callbacks for interviews.

First, we used the aforementioned within-subjects experimental design, with Boston RAs randomly assigned to travel using either a "white" or "African American" profile. Second, we instructed RAs to carefully watch for, and systematically record, trip cancellations.

Discrimination by TNC drivers may occur in at least four primary ways, highlighted in Figure 2: (i) drivers could elect not to drive in or near certain types of neighborhoods, (ii) drivers could decline to accept reservations from certain types of passengers or could cancel a pickup once the passenger's identity becomes revealed, (iii) drivers could leave low ratings for passengers based on race, gender, or perceived socioeconomic status, and (iv) drivers could choose non-ideal routes based on the same factors, increasing costs and/or travel time. Our analysis focuses on the second, fourth, and, to a lesser extent, the third of these channels.

Our results fit within existing work that has demonstrated evidence of profound discrimination in the peer economy.[2] In a non-experimental setting, Pope and Sydnor (2011) showed that loan requests were significantly less likely to be successful when associated with an African American person than when associated with a white person of comparable credit history. Discrimination can occur even without seeing a person. Ongoing work by researchers at Harvard Business School has demonstrated racial discrimination on both sides of a peer-to-peer transaction on the Airbnb platform (Edelman and Luca, 2014; Edelman et al., 2016).

Our work also fits within an existing literature that has identified discrimination in the transportation sector. For example, Goddard et al. (2015) finds discrimination behavior of drivers toward pedestrians of different races, with African American pedestrians at a crosswalk being passed by twice as many cars and waiting 30% longer for a car to stop than white pedestrians. Other research has shown that perceived differences in social class affected vehicle-vehicle and vehicle-pedestrian yielding behavior (Piff et al., 2012). Given the presence of discrimination in more traditional parts of the transportation network, we do not claim that TNC networks are "worse" than the status quo. In fact, in the Seattle pilot we had RAs hail taxis and count the the number of empty taxis passing them by. There is a clear difference in the acceptance rate of traditional taxis. The first taxi stopped nearly 60% of the time for white RAs, but less than 20% of the time for African American RAs. The white RAs never had more than four taxis pass them before one stopped, but the African American RAs watched six or seven taxis pass them by in 20% of cases.

Our work explores options for reducing discrimination within TNCs. For example, TNC networks could omit personal information about potential riders completely. Names and photos of both passengers and drivers could be replaced with user-specific numbers. Confirmation of these numbers by the driver and passenger could then occur prior to the trip. While this would reduce the type of discrimination found in this paper, other channels of discrimination would remain. Most notably, it would not eliminate the ability of drivers to

[2]Research has also found evidence of discrimination in online-sales markets based on the skin color of the person holding the item for sale. See, for example, Doleac and Stein (2013) and Ayres et al. (2015).

discriminate in where they choose to drive.

The paper unfolds as follows: In section 2 we describe the design of our Boston experiment, the results of which we present in section 3. In section 4 we return to the Seattle pilot, describing the design of and results, as well as how it informed our Boston experiment. Section 5 concludes the paper.

2 Experimental Design

The Boston experiment used a within-subjects design, with eight research assistants (RAs) each traveling under two assumed profiles. RAs were issued two identical phones, each with Uber and Lyft applications installed and with a travel profile under one of the assigned pseudonyms. To reduce the likelihood that RAs behaved differently under one profile or another, neither pseudonym was related to the traveler's true name. This had the additional benefit of preserving the travelers' anonymity for the duration of the project.

2.1 Data Collection

Eight individuals traveled as part of the experiment in Boston, using the pseudonyms shown in Table 1. These pseudonyms were taken from lists of names developed by Bertrand and Mullainathan (2004) that had been strongly identified as African American or white by panels of observers. Paying for the rides presented a challenge, since travel profiles were not real individuals and did not have credit cards. Payment was initially made using institutional purchase cards, but rapid accumulation of Uber and Lyft transactions on a few credit cards raised fraud flags by issuing banks, Uber, and Lyft. To avoid account suspension and potentially stranding travelers in remote parts of the city, we extended payment methods to include their personal credit cards, and eventually PayPal accounts under the pseudonyms issued to the traveling RAs. However, even this became a challenge as both the accounts under the pseudonyms were cancelled by Uber. Ultimately, these challenges led the experiment to be terminated slightly prior to the planned termination date. The experiment ran from November 5, 2015 to March 3, 2016.

2.2 Site and Route Choices

We developed specific travel tours that consisted of a list of stops with specific pickup and drop-off points. These tours ensured that we could control for the demographics of the pickup location and the expected distance and duration of each trip. We performed a GIS analysis of the Boston area, shown in Figure A.4. In determining which location attributes to include in the analysis, we hypothesized that drivers would accept or reject rides based on a mixture of fear and greed. For example "What is my expected revenue for this and future rides, and what is my fear that harm will be done to me by this passenger or others in the pickup location?" We examined pickup locations for household income, population density, measures of minority population density, and the presence of a transit stop at the point of pickup. We used the demographics at each location to help design tours as described below.

The final research design used five unique tours constructed with multiple objectives in mind: (i) start and end near the same place, (ii) reduce overall cost by limiting the distance and duration of each trip to near the minimum fare threshold, (iii) induce as much variation in pickup location demographics as possible, (iv) minimize overlap with other tours to reduce likelihood of multiple travelers affecting performance for each other, and (v) where possible, to ensure pickup locations were within sight of a business open until at least 9 pm.

The RAs alternated between requesting UberX and Lyft rides, with the first mode of each tour assigned randomly. Rather than indicating the pickup location on a map, RAs were instructed to enter the complete address of the pickup location and destination to prevent small variations in pickup location based on individuals. Since RAs were traveling late in the day and sometimes in bad weather, in a few instances RAs requested to end the last trip of the tour at their home address. These trips, and trips for which RAs realized they had entered the origin or destination address incorrectly, were excluded from trip distance and duration analyses.[3]

2.3 Traveler Instructions

While the empirical analysis includes RA fixed effects to control for individual variations in behavior (using the Uber and Lyft apps or at the point of pickup), RA travelers followed specific instructions in requesting and taking rides to minimize differences across RAs. These instructions were: (i) to enter the complete address of both the pickup location and the destination when requesting a ride to minimize variation in origins and destinations, (ii) to minimize interaction with the driver to lower the chances that a driver would recognize them, and minimize any differences in behavior that would change star ratings, (iii) not to cancel rides unless a ride was input incorrectly, (iv) not to give drivers directions unless requested, and (v) to leave five-star ratings unless they felt the driver was threatening or dangerous.

The RAs watched continuously for driver cancellations and logged key information by taking screenshots on their smartphones. We installed an app on each smartphone that displayed the time including seconds, so we could easily read the precise time in each screenshot. For each trip, we instructed them to take the following screenshots:

[3]We were aided by the Seattle pilot (see section 4) for calculating statistical power. We calculated the sample size required to detect a variety effects on the means of the variables described above using the summary statistics from Seattle. We calculate the samples sizes required to detect a 5, 10, and 15 percent increase in the variables of interest. We used the UberX sample for our calculations and assume an α of 0.05 and power equal to 0.90. These calculations implied sample sizes of roughly 1,700, 400, and 200 for detectable effects with actual wait times of 5, 10, and 15 percent, respectively. The full results are reported in Appendix Table A.4. To err on the conservative side, we performed these calculations without including any covariates. In the next section we describe what covariates we included to gain some statistical power. Based on these calculations we targeted 1,000 data points for the Boston experiment. Because of the early termination of the data collection in Boston we ended with 911 trip requests.

1. Immediately before requesting a trip. This captures the time when a trip was requested, and the estimated waiting time for the passenger to be picked up (displayed in the TNC app).

2. Immediately after a trip request was accepted by a driver. This captures the time when the trip was accepted and a revised estimated waiting time for the passenger to be picked up.

3. Immediately when a driver cancels a trip.

4. Immediately upon a new driver accepting after another driver has cancelled.

5. When the driver arrives to pick up the RA. This provides the actual pickup time.

6. When the car stops to drop the RA off at the requested destination. This captures the actual dropoff time.

The RAs took notes on additional relevant information that arose in the course of their tours, such as deviations from the prescribed experimental plan, problems with data collection, or practical challenges with the prescribed stop locations.

The RAs transcribed key data from their smartphones into a spreadsheet at the end of each tour. The screenshots were deleted from the smartphones after transcription because some of them contained personally identifiable information (names and photos) of the drivers. In addition, we obtained the price and distance traveled for each trip from receipts that were automatically emailed to the research team at the end of each trip. We also deleted the emails containing driver names and photos.

After several initial travel days looking for cancellations, RAs identified an unanticipated form of driver behavior. Some drivers would accept a ride and then apparently not attempt to pickup the passenger: not moving at all, beginning a ride without the passenger, or driving in the direction opposite the traveler. Since this scenario could result in RAs waiting indefinitely for a ride in unfamiliar locations, we added an instruction. If the driver started a ride without the passenger or had not made any indication of attempting to pickup the passenger after 15 minutes—either contacting the passenger or driving measurably closer to the pickup location—RAs cancelled the ride and flagged the first attempt as a *de facto* cancellation by the driver.

3 Empirical Strategy and Analysis

In our experiment, treatment is whether the rider has what Fryer and Levitt (2004) refer to as a "distinctively black name." Since each of our RAs used two pseudonyms we were able to include RA-specific fixed effects in our econometric specifications. The starting point for our empirical analysis is to estimate differences in the within-RA means of the variables described above across implied race. More specifically, for a given outcome y (e.g., acceptance time, waiting time, etc.), for RA i, using name k, at location j, at time t, our regressions take the form of:

$$y_{ikjt} = \beta_0 + \beta_1 \text{ImpliedRace}_{ik} + \mu_i + \epsilon_{ikjt}. \tag{3.1}$$

We will also estimate specifications that control for covariates.[4] These covariates capture location, trip, and day-of-week effects. We therefore augment Equation (3.1), yielding:

$$y_{ijkt} \;\; = \;\; \beta_0 + \beta_1 \text{ImpliedRace}_{ik} + \gamma \mathbf{X_j} + \mu_i + \epsilon_{ikjt}. \tag{3.2}$$

We rely on three measures of quality of service as dependent variables. They are:

- **Acceptance time** is the length of time that passed between when an RA sent a trip request and when that request was accepted by a driver.
- **Actual waiting time** is the length of time that passed between when a trip was accepted and when the driver arrived to pick up the RA.
- **Cancellation** is an indicator variable for whether the initial request for a ride was accepted by a driver and then subsequently cancelled.

We might expect to see differences in *acceptance time* between races if drivers discriminate. In the case of Lyft, which showed the passenger's profile photo to the driver as part of the trip request, a driver's likelihood of declining the request might be affected by the traveler's apparent race. In the case of Uber, which showed the passenger's first name to the driver *after* the trip is accepted, this seems less likely. However, it is possible that some drivers might accept and then quickly cancel a trip after seeing the passenger's name, behavior that some drivers have advocated in online forums (UberPeople.Net, 2015). In either case, the request would then be passed to another driver, which would lead to a longer delay between the time the passenger requested a trip and when that trip was ultimately accepted.

We might also see differences in *actual waiting time* if drivers discriminate. Let us first assume that a trip request is reliably routed to the nearest driver and then to progressively further drivers if the closer driver(s) decline (or accept then cancel) the request. If at least some drivers tend to decline requests from certain groups, then on average those groups would be matched with drivers who are farther away, and would end up waiting longer for a car to pick them up. On the other hand, if the matching of requests to drivers depends on variables other than distance, this effect would be attenuated, since declined requests would

[4]In particular we include indicator variables for whether the pickup location has a high African American population, a high median income, and a low population density. Demographics were assigned by Census Block Group of the pickup point using 2010 data. Upper AA Pop Quartile is defined as 20% or greater African American population, Upper Income Quartile is defined as median household income greater than $60k annually, and Lower Population Density Quartile is defined as fewer than 12800 people per square mile.

sometimes be rerouted to drivers who are actually closer than the original driver. We would expect the effect to be strongest for actual wait times, which is our focus below.

As noted, discrimination may also manifest itself through cancellations. Here, we would only expect to see this for Uber rides given that Uber drivers do not see the passenger's name until they have accepted the ride. Therefore, if drivers infer race from passenger names they would reject the ride only after first accepting it. In the case of Lyft, by contrast, drivers see the name prior to accepting the ride and can simply ignore a prospective passenger's request.

3.1 Results

We collected data on 911 total trip requests in Boston: 460 trips with Lyft and 451 trips with UberX. Drivers cancelled 66 requests (57 actual cancellations, and 9 *de facto* cancellations in which the driver started the trip without the RA or made no apparent pickup attempt). RAs cancelled 6 trips after recognizing they made input errors in the requests. After excluding cancellations, 839 completed trips remained.

Out of the 839 completed trips, we removed from our analysis of acceptance and waiting times all observations in which travelers failed to capture a screenshot of one or more timestamps during the trip. 80 trips were removed, leaving 759 trips in which all valid timestamps were observed—372 using UberX, and 387 using Lyft. Table 2 provides summary statistics for request times, wait times, and estimated waiting times for UberX and Lyft in Boston. In cases where a driver cancellation occurred, RAs recorded the time of the initial trip request as the request time. That is, the acceptance and wait times include the added time generated due to the cancellations. As a result, both Lyft and UberX show large standard deviations in acceptance times, and median acceptance times are far lower than means for both services.

We begin by confirming that we have balance across key trip descriptors. Table 3 presents the means of *estimated* wait times, surge pricing, and three pickup-location features when RAs used the African American and white-sounding names, as well as the difference in the means. Standard errors are reported in parentheses. Estimated wait time is the waiting time that was displayed in the app immediately before the RA requested a ride, and therefore should not differ across the name used. None of the variables systematically differed across the names. Furthermore, the differences are small relative to the means with the possible exception of low-population-density locations which differed across populations by roughly 13 percent.

Table 4 shows the results of regressions for acceptance time in Boston, and Table 5 shows the results of actual waiting time in the Boston experiment. We do not find a statistically significant increase in wait times when RAs used the African American names. While all positive, none of the coefficients are statistically significant in our acceptance time specifications. The point estimates for acceptance times range from 1.3 percent (Model 6) to 4.1 percent (Model 2). The upper bound of the confidence intervals is an increase in acceptance times of 20 percent (Model 2), or roughly 10 seconds. For wait times, the point estimates range from a reduction of 7.8 percent (Model 1) to a reduction of 0.5 percent (Model 6). We can rule out an increase in wait times of larger than 12 percent (Model 6), or 35 seconds.

The heart of our analysis is in cancellations. Since UberX drivers only see the name of passengers after accepting a ride, we expect to observe differences in behavior only in UberX. On the other hand, since Lyft drivers see passenger names before accepting rides, if they were to discriminate based on name, we would expect them to simply ignore the ride request, an action which is not directly observable in our empirical setting.

We first tabulate the rate of cancellations of trips by service, race, and gender. This is shown in Table 6. The simple summary statistics suggest that cancellations against passengers using African American names are substantially higher for trips with UberX, but not for Lyft. For all passengers on UberX, those using African American-sounding names face more than double the cancellation rate than when the same passengers use white-sounding names (10.1% vs. 4.9%). Furthermore, this effect appears larger for African American male names than for African American female names.[5]

When using UberX, the race effect on cancellations is greater for males than females. For males, those using African American-sounding names face a cancellation rate more than twice as high as those same individuals when using white-sounding names–11.2% cancellation rates for males when using African American names compared to 4.5% when the same individuals use white-sounding names. The difference is smaller for female riders: the cancellation rate for females is 8.4% when using an African American name and 5.4% when using a white-sounding name. However, for Lyft, males face approximately the same cancellation rate, whereas females actually face a lower cancellation rate when using African American names.

Next, we use regression to gauge statistical significance and to control for a broad array of demographic variables for the pickup locations, individual fixed effects for the passengers, and fixed effects for the day of the week. The results from these regressions are shown in Table 7. We start by analyzing only the means. For Lyft, we find no statistically or economically significant difference across names. The treatment effect for UberX, however, is indeed significant and suggests that the cancellation probability for riders using African American names more than doubles, consistent with the cross-tabs above.[6]

We expand the analysis along two further dimensions. First, we include a variety of controls, such as demographics of the pickup location and the day of week. Second, we interact the treatment with gender. For completeness we allow the white baseline cancellation probability to also vary by gender. Column 2 in Table 7 presents these results for Lyft. Again, we find little evidence of an increase in cancellations across race or gender. This again serves as a useful null test since drivers can observe the prospective passenger's name prior to

[5]We directly test for the significance of this difference in the regressions below.

[6]There may be correlation in the errors within RA given the nature of the experiment. With so few RAs, clustering at the RA level is not practical. Instead, we have also calculated standard errors using randomized inference using the ritest command in Stata stratifying at the RA level. In each case, the standard error shrinks.

accepting. Column 4 reports the results for UberX. We find that discrimination appears to be focused on African American males. The cancellation rate faced by travelers using a stereotypically African American male name more than doubles (139% increase) relative to a stereotypically white-sounding male name. To gauge the robustness of this result to RA-specific behavior, Column 5 includes RA fixed effects. Column 6 adds an indicator variable for whether the pickup location is at a subway station. The estimate on the African American male indicator is robust to the inclusion of these.[7]

We also investigated whether discrimination was isolated to specific pick-up areas and time periods (columns 7 & 8). We interact the African American male indicator variable with the low population density dummy variable. Interestingly, we find that the increase in cancellations is concentrated in areas with low population density. In fact, the cancellation rate of riders with African American names is no different outside of these areas. But within these areas, males with African American names face an increase in the cancellation rate of 16.0 percentage points; more than three times that faced by white males.

Finally, we test whether surge-price events affect discrimination in Column 8. We interact the African American indicator variable with the surge multiplier, which ranges from 1 (standard pricing) to 2.6. The point estimates suggest that discrimination is more likely during surge pricing, although the coefficient is only marginally statistically significant. Including this interaction term does not alter the results with respect to the low population density interaction term. We suspect that when surge pricing is active, there are two countervailing effects: (1) the cost of declining a ride is higher since that ride would have brought a higher fare, and they may miss out on this higher fare if they wait; and (2) drivers expect to get rematched with a new passenger more quickly, since surge pricing occurs when there are fewer drivers relative to passengers. The point estimates suggest that this second effect dominates, as discrimination is more likely during surge events.[8]

Admittedly, it is somewhat of a puzzle why we do not observe a statistically significant increase in wait times, even though African Americans are twice as likely to be canceled upon. As discussed in the following section, the Seattle pilot suggested that African American RAs faced longer wait times, although these effects were not statistically significant.

We also explored the impact of driver density within the Boston data, leveraging the fact that driver density is likely lower in low-population-density areas. First, we note that the average acceptance and wait times are longer in areas with low population density, at 67.5 and 333.7 seconds, respectively, compared to 49.6 and 300.0 seconds elsewhere. This implies that not only is the increase in cancellation rates greater in these areas for African American males, the effect this will have on wait times is also larger. We directly test whether this is

[7] Please note that we are unable to reject equality between the female African American coefficients and either the male African American or the female white coefficients. The p-value in Column 2 of the test between African American women and white women could be considered marginally significant at 0.091.

[8] The point estimate for the surge event interaction term changes very little if we omit the low population density interaction term. When this latter interaction is omitted the results suggest that the probability of a cancellation for an African American male increases by 5.96 percentage points; the p-value for this is 0.0977.

the case within the acceptance time and waiting time regression models estimated above by adding a variable that interacts an African American male indicator variable with the low population density indicator variable, focusing on the UberX rides. Table 8 shows the results of this exercise for acceptance and wait times. Here, we see that African American males faced both higher acceptance and total times when requesting a ride in a low-population-density area. The point estimate, 0.31, is statistically significant for acceptance times; the sum of the African American and African American low population density interaction term equals 0.27 and is marginally significant ($p = 0.07$). For wait times the point estimate is positive, but not statistically significant, and the 95% confidence interval for the interaction term ranges from -0.09 to 0.38. The confidence interval for the combined effect of the African American indicator with the interaction term of 0.06 ranges from -0.16 to 0.28. These results suggest that when we combine the increased propensity of cancellations with the lower driver density of these areas, there may be measurable impacts on wait times.

It would be wrong (if tempting) to conclude that since we did not detect statistically significant differences in waiting times, TNC discrimination does not lead to economic costs. Such an interpretation would take too narrow a view of economic costs. For one, while the confidence intervals are wide (and with the exception acceptance times among African Americans in low population density areas, contain zero) it is important to note that large wait time effects are also possible. Second, the distribution of the cost of being late is likely to be highly skewed; for example, being late to the airport often has zero costs, but can on rare occasions have tremendous costs. While we are underpowered to make precise statements, we also estimated quantile regressions of acceptance and wait times for the Boston experiment. If delays in wait times were coming from cancellations, we might expect the effect of African American names to be larger for larger quantiles of the residual. Appendix Figures A.5 and A.6 plot the quantile regression coefficients associated with the African American indicator variable varying quantiles by 0.05 from 0.05 to 0.95. Also included in these figures are the quantile 95% confidence intervals and the OLS estimate. We see some evidence that the estimated effect of race increases as the quantile increases, but the evidence is not very strong. If an African American passenger were unlucky enough to be faced with one of these large delays at a time where he or she is late for a flight, or a job interview, the economic costs could be large, albeit difficult to see in the means. Finally, we note that there are likely to be important psychological costs associated with discrimination. Our experiment is not designed to measure such costs, but anecdotal evidence suggests that society, rightly so, generally takes a negative view of discrimination even when it leads to separate, but equal, service quality.[9]

[9] We also investigated how trip distance and trip time correlated with gender and race. If discrimination manifests itself as a small subset of drivers discriminating, as opposed to the majority of drivers discriminating a "little," the effect of discrimination is likely to show up at the high end of the distribution of acceptance and wait times. These results are reported in Ge et al. (2016). There is some evidence that travelers are driven farther. We freely admit that we did not design the experiment to directly test this; we encourage future research in this area. OLS regressions using the log of travel distance indicates that female travelers are driven approximately 5% farther. We repeated the quantile exercise discussed above for gender. Appendix

4 Seattle Pilot

Prior to our field experiment in Boston, we ran a pilot experiment in Seattle, WA. This pilot study served to familiarize ourselves with the Uber and Lyft platforms, helped us manage research assistants, and provided data for power calculations. We describe the pilot here so the reader may more fully understand the context and motivations for the design of our Boston experiment. Instead of randomizing within each RA, African American and white research assistants (RAs) used UberX, Lyft, Flywheel (app-based taxi hailing), and taxis hailed from the curb to traverse assigned routes within the city of Seattle over six weeks in August and September, 2015. We tested for differences between races in the speed of service.

We designed seven tours around the city of Seattle, each starting and ending at the University of Washington's Seattle campus and comprising a sequence of pre-determined stop locations linked by individual trips. The stops were located to generate variability in neighborhood characteristics (population density, percentage of residents who are African American, income level), while limiting the individual trips to roughly the distance corresponding to the UberX and Lyft minimum fares. The routes are mapped in Appendix Figure A.1, which also shows selected socioeconomic characteristics at the census block group level. The tours generally took between one and three hours and were completed following the evening rush hours on Monday through Thursday evenings. To avoid confounding the effects of race, sex, and other variables, we generated a fractional factorial experimental design. The variables and levels used in the experimental design are summarized in Table A.1. In this way, we produced a list of tours to be completed on specific days of the week, by travelers of a particular race and sex, beginning with a specified service.

In the first four weeks of the study, we assigned the RAs to rotate among services in the general order UberX-Lyft-Flywheel, after starting each tour with a randomly specified service. At selected stops in downtown Seattle, they were directed to hail a passing taxi from the curb (hailing a taxi from the curb is not feasible in other areas of Seattle, due to the low density of taxis). In the final two weeks of the study, we stopped collecting data using Flywheel, and the RAs alternated between UberX and Lyft (while still hailing taxis from the curb at specified downtown stops).

We began data collection with eight RAs: two African American females, two African American males, two white females, and two white males. All of the RAs were University of Washington undergraduate students who were recruited through an on-campus job posting

Figure A.7 plots the Quantile regression coefficients associated with the female indicator variable. While the coefficient is positive for each of the quantiles, the coefficient appears to increase after the median. This behavior was confirmed by anecdotal evidence from female passengers. Some excessive fares were the result of drivers who started the trip before picking up the passenger or ended the trip after dropping off the female passenger. Other female riders reported "chatty" drivers who drove extremely long routes, on some occasions even driving through the same intersection multiple times. As a result, the additional travel that female riders may be exposed to appears to be a combination of profiteering and flirting to a captive audience. This is important because TNCs have moved to compensation algorithms based only on the distance between the origination and pickup points, and not the actual driving distance or driving time.

board and were paid $17 per hour. We presented the RAs with a list of dates on which the experimental design dictated that a traveler of their race and sex should travel, and they signed up for specific travel days. Each RA completed no more than one tour in a day and used smartphones to request rides from UberX, Lyft, and Flywheel, and to log data. We issued each RA an identical smartphone using the same mobile carrier and data plan to minimize variation in factors such as communication latency. The RAs set up passenger accounts with Uber, Lyft, and Flywheel on these smartphones, and included their actual first name and a profile photo with each account (Flywheel did not support profile photos). Profile photos were taken during the RA training session, and consisted of a headshot of each RA with a neutral facial expression, in front of a plain white background. The first names of the white RAs were Elizabeth, Rachel, Mitchell, and Sergey; those of the African American RAs were Reynide, Surita, Malik, and Adetimi.

The RAs logged the same information as described in the Boston experiment, except we did not instruct them to collect data on cancellations.[10] For taxis hailed from the curb, the RAs took screenshots when they began trying to hail a cab and when a cab stopped for them. They also kept a count of how many taxis passed by them before one stopped and took a note of this on their phone. To avoid difficulties in identifying which taxis were available and which were already serving other passengers, RAs counted *all* taxis that passed them while they were trying to hail.

To minimize differences in how the RAs conducted the data collection, we conducted a two-hour group training session the week before data collection began. The RAs were introduced to the project and familiarized themselves with the smartphones and the Uber, Lyft, and Flywheel apps. We provided instructions on the data collection process and on how to interact with drivers. As with the Boston experiment, we instructed the RAs to sit in the back seat, minimize their interactions with the drivers, and not indicate that they were collecting data. We also instructed them to enter their destination information via the apps after the driver arrived to pick them up, and if asked, to request that the driver simply follow the navigation app linked to their TNC platform.[11]

[10] Examples of the screenshots are presented in Appendix Figure A.2.

[11] There were nevertheless some sources of variability and deviations from the original experimental design. First, due to the RAs' scheduling constraints, we allowed the them to complete some tours on different days than originally prescribed (e.g., a day earlier or later), but all the trips were completed between Monday and Thursday. Second, we asked the RAs to begin their tours at approximately 6:30 pm each day, but allowed them flexibility to start 30 minutes before or after this time. Third, we gave the RAs some latitude to adapt their data collection on the fly. For example, if dynamic pricing ("surge pricing" for Uber or "primetime" for Lyft) increased TNC fares to more than 2.5 times their base rate, we asked the RAs to switch to another service for that trip. The RAs also switched modes if they faced repeated cancellations or were otherwise unable to complete a trip by the prescribed service. Finally, approximately three weeks into the study, one of the African American male RAs was unable to continue with the project. We hired another African American male RA to take over his smartphone and TNC accounts (after replacing the profile photo with a photo of the new RA) and complete his remaining trips.

4.1 Pilot Empirical Strategy and Results

We used linear regression to test for differences in acceptance and wait time across African American and white RAs. The general form of our regression model is similar to our Boston experiment specifications and is shown in Equation 4.1, for some outcome variable y (e.g., acceptance time, wait time, etc.) for RA i at location j in situation t. Race_i is a dummy variable indicating RA i's race, and $\mathbf{X_{ijt}}$ is a vector of covariates unique to the individual, location, and/or situation. For example, $\mathbf{X_{ijt}}$ can include neighborhood characteristics or the estimated waiting time associated with the trip request.[12] This yields:

$$y_{ijt} = \beta_0 + \beta_1 \text{Race}_i + \gamma \mathbf{X}_{ijt} + \epsilon_{ijt}. \tag{4.1}$$

4.2 Pilot Results

We collected data on 581 app-hailed trips in total: 208 Uber trips, 222 Lyft trips and 143 Flywheel trips. Among these trips, 155 were finished by African American female RAs, 152 by white female RAs, 129 by African American males, and 145 by white males.[13] The fact that fewer trips were taken by African American males compared to other groups is due to the aforementioned replacement of one African American male RA with another halfway through the data collection process.[14]

Table A.2 provides summary statistics for acceptance time, estimated waiting times, and actual wait times of Uber, Lyft and Flywheel. The number of observations for the acceptance time and actual waiting time of the three modes are different because some observations were deleted for the following reasons: (1) the RAs did not take a screenshot immediately before requesting a trip, immediately after a trip request was accepted, or immediately when the driver arrived; (2) the times on the screenshots are missing or inaccurate because the clock was fully or partly blocked on the phone; or (3) the RAs made typographical errors during data entry (e.g., the time right before requesting a trip is later than the time right after the request was accepted). Some observations of the estimated waiting time 1 were deleted because the estimated waiting time did not show up when the service was busy. Comparing the three services, Lyft requests are accepted more quickly, on average, but Uber has the

[12] In particular, we include demographics from the Census Block Group of the pickup point using 2010 data. We include three indicator variables. "High Income" is defined as median household income greater than \$75k annually (top 25%), "High Pop Density" is defined as fewer than 6750 people per square mile (lower 25%), and "High AA Pop" is defined as more than 1890 African American people per square mile (top 25%).

[13] We also collected data on 36 taxi trips in which the taxi was hailed from the curb. We come back to this at the end of the paper.

[14] Appendix Figure A.3 shows the balance in the days, times, and locations (tours) of the experiments actually done by the African American and white RAs. The distributions of the day of week and time of day are similar between the races, but white males completed more trips on tour 4 than African American males. Since tour 4 goes through downtown Seattle and the Capitol Hill neighborhood, where supply and demand may be different than elsewhere in the city, this imbalance in routes could lead to differences in waiting times.

shortest average waiting time. Flywheel is the slowest when it comes to both trip acceptance time and waiting time.

We first test whether there are systematic differences in the *estimated* wait times reported by the services' apps. These results are presented in Appendix Table A.3. The results indicate that before trips were requested, there were no significant differences in the estimated waiting times presented to African American and white RAs.[15]

Table 9 presents the parameter estimates from our full regression model that includes route fixed effects, controls for estimated wait time, and includes covariates capturing income, population density, and the proportion of African American population at the pickup location. Each row is a separate regression across the log of the measure of quality (acceptance time and actual waiting time) and service (Uber, Lyft, and Flywheel). Inference needs to account for the fact that randomization occurred between the research assistants, rather than within the research assistants as in the Boston experiment. To account for this, the reported p-values are calculated using the randomization test in Young (2018). In particular, we first run the regression and capture the estimated African American effect. We then run a Monte Carlo where for each iteration, RAs are randomly assigned race categories and the regression is re-estimated. The p-values are then calculated as the percentage of time the African American coefficient is larger than the non-RA randomized coefficient.

The p-values imply that although the estimated effects are positive in five of the six specifications, the parameter estimates are statistically insignificant. Therefore, while the results are consistent with discrimination, they lack statistical power, driven by the smaller sample size and the between-RA design, to draw any conclusions.

4.3 Lessons Learned

Our pilot revealed a number of potential limitations to the research design, and we applied the lessons learned in Seattle to refine the experimental design for the Boston study. For one, it is conceivable that differences in measured acceptance times or waiting times might be due to systematic differences in how individual RAs logged their data, and this was somehow correlated with race. Although all of the RAs underwent the same data collection training, it remains possible that the differences in acceptance time might be due to differences in data logging practices between the African American and white RAs. For example, if African American RAs had a longer lag between taking their initial screenshot and sending the trip request, or having a trip accepted and taking their second screenshot, this would lead to longer measured acceptance times for trip requests by the African American RAs, even if the actual time from sending the request to having it accepted were the same for both groups. We doubt this is the case, since we would expect to see this consistently across all platforms, yet we did not see any difference in acceptance times between African American and white

[15]Although the difference is not significant, the estimated waiting times for African American travelers were slightly longer than for white travelers, which may be related to the aforementioned imbalance in the number of trips taken along tour 4.

travelers when using Flywheel. Moreover, this cannot explain the larger difference in average waiting time (roughly 90 seconds) observed between African American and white passengers using UberX.

A second limitation is that it is possible that differences between African American and white passengers using UberX were due to some drivers having trouble identifying the African American passengers at the pickup points. If, for example, drivers were not expecting an African American passenger, then it might take them longer to see the passenger and drive up to them. This could explain why the pickup times were longer for African American passengers on UberX, even if there were no overt discrimination; although the driver arrives for the pickup in the same amount of time, they might spend more time looking around for the passenger. This might also explain why the travel times for UberX were longer, although the travel distances were not; UberX might record the trip as starting when the driver arrives at the pickup location, even if drivers sometimes spend a little extra time looking around for their passengers. This could also explain why these effects were not detected on Lyft: the drivers have a photo of the passenger from the outset so they know exactly who they are looking for.

Finally, we did not design the Seattle pilot to understand the precise mechanism for discrimination by drivers receiving an UberX trip. As noted, drivers do not receive any information about the passenger until after they accept the request, so it would seem that in order to discriminate they would need to cancel an accepted trip. Since we did not originally anticipate the possibility of drivers accepting then canceling trips, we did not provide the Seattle RAs with clear instructions about logging cancellations. Sometimes the RAs noted that a cancellation had occurred, but we are not confident that they did so in all cases. It is possible that a driver could cancel and UberX could assign a new driver, without the RA noticing.

Ultimately, the lessons learned during the Seattle pilot led us to make two major changes to the experimental design for the Boston study. The first is that we designed our study in Boston to use within-RA variation in implied race to control for differences in data collection practices across RAs. Furthermore, we recruited RAs with a range of ethnic backgrounds, whose appearance would allow them to travel under either of their assigned names without attracting attention. The second change was that we instructed the RAs to watch vigilantly for cancellations. As noted above, there is active discussion on driver forums (e.g. UberPeople.Net, 2015) about whether cancellations that are performed quickly are shown to a customer. If drivers can cancel quickly and not appear on a customer's screen, then measurements of cancellations by RAs should be treated as a lower bound, and actual cancellations could be higher than those reported. A third and less substantive change was that due to the increased focus on cancellations, we turned our focus to the largest TNC services UberX and Lyft, and did not perform tests of FlyWheel or street hails in Boston.

5 Conclusions

The digital platforms of transportation network companies such as UberX and Lyft could help to limit discrimination because they can control what drivers know about the passenger prior to pickup, or could perpetuate historic discrimination by taxi drivers. To better understand the degree of discrimination present in current ride-sourcing networks, we designed a randomized field experiment in Boston where we varied the perceived race of passengers by having RAs randomly choose between two first names. This experiment was informed by a smaller between-subjects pilot in Seattle.

The pilot results suggested that African American travelers in Seattle experienced longer delays waiting for a trip request through UberX or Lyft to be accepted, though the difference was not statistically significant. For total wait times, we found larger African American effects when RAs used UberX, compared to when they used the Lyft service. Again, the results were suggestive but not statistically significant. Nevertheless, we noted that UberX drivers see only a passenger's location and star rating before accepting or declining a trip request, and see the passenger's name after accepting. Lyft drivers can see the passenger's name and photo before accepting or declining the request. Therefore, discrimination requires an additional step with the Uber service; drivers have to first accept a ride and then cancel.

Given these results, we designed the Boston experiment to more closely focus on cancellations. We find that UberX drivers are nearly three times as likely to cancel a ride on a male passenger upon seeing that he has a stereotypically African American name. This effect is robust across numerous model specifications and seems to be driven primarily by behavior in areas with low population densities. In these extreme cases drivers are more than four times as likely to cancel on an African American male passenger than on a white male passenger. As expected, given that Lyft drivers see passenger information prior to accepting a ride, we find no effect of race on Lyft cancellations. We also find that UberX drivers are more likely to cancel trips for passengers being picked up near subway stops, perhaps because a passenger at a subway stop is either a low-income passenger, or because a subway stop indicates a multi-modal journey with a lower expected revenue.

Using the most direct measure (observed cancellations in Boston), there appears to be evidence that African American passengers receive worse service, compared to white riders, in TNC services such as Uber and Lyft. There is no reason to believe that this discrimination is the result of any policy by ride hailing providers, and we believe it emerges from the behavior of individual TNC drivers. It is important to note, however, that we compared service quality across African American and white passengers *within TNC services*. That is, we do not compare the relative amount of discrimination across TNC and traditional taxi-cab services. In our Seattle pilot we found that discrimination within traditional taxi-cabs also exists. The first taxi encountered stopped nearly 60% of the time for white RAs, but less than 20% of the time for African American RAs (Figure A.10). Furthermore, white RAs never had more than four taxis pass them before one stopped, but the African American RAs watched six or seven taxis pass them by in 20% of cases. These differences are statistically significant. Since it is practical in Seattle to hail a taxi from the curb only on a few downtown streets, the data collected in this way are not directly comparable with the TNC or Flywheel data that were collected city-wide, but the curb-hailing data do indicate that discrimination is not restricted to Uber and Lyft drivers.

This study was not designed to distinguish between statistical and taste-based discrimination, but the "star ratings" that the drivers provided for the research assistants provide some insight into this question. As shown in Figures A.11, A.12, A.13 and A.14, drivers provided the African American and white RAs with virtually identical star ratings, on average. To explain this via statistical discrimination, discriminating drivers would have to believe there is some important measure of passenger quality associated with race, *beyond* the factors that other drivers are including when assigning their star ratings. For example, discriminating drivers would have to believe not only that accepting an African American passenger means a longer deadhead after the trip, but that other drivers have not considered this when assigning their star ratings. Thus, we conclude that our data point toward taste-based discrimination.

6 Tables and Figures

6.1 Figures

Figure 1: TNC ad featuring an African American rider

Figure 2: Potential sources of discrimination

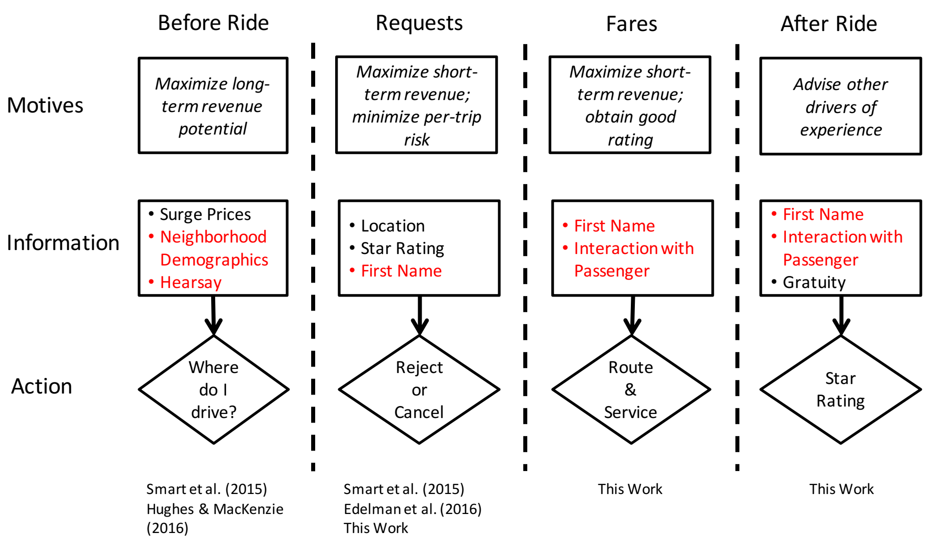


Table 1: Subject name and race assignment in the Boston experiment

RA	Gender	African American Name	White Name
1	F	Aisha	Allison
2	F	Ebony	Kristen
3	M	Hakim	Brendan
4	M	Darnell	Brad
5	F	Keisha	Anne
6	M	Kareem	Greg
7	M	Rasheed	Todd
8	F	Latoya	Laurie

Research assistant names used in the Boston experiment. The first name is meant to reflect a stereotypical African American name, while the second is meant to project the research assistant as white. These are drawn from the names used in Bertrand and Mullainathan (2004).

6.2 Tables

Table 2: Descriptive analysis of the acceptance time, estimated waiting time 1 and actual waiting time of Uber and Lyft in the Boston experiment

	Attributes	Acceptance Times	Est. waiting time 1 (min)	Actual waiting time (s)
Uber	Mean	51	3.3	218
	S.D.	111	1.4	148
	Median	25	3	196
	African American Mean	51.9	3.3	211
	White Mean	50.7	3.3	226
	Num.obs	372	372	372
Lyft	Mean	54.6	2.8	297
	S.D.	172	1.4	211
	Median	19	2	243
	African American Mean	55	2.7	296
	White Mean	54	2.8	299
	Num.obs	387	387	387

Notes: This table shows the number of trips taken and summary statistics for the time until the request was accepted (Acceptance time), the *estimated* wait time reported by the app when the trip was requested (est. wait time 1), and the actual wait time from when the trip was requested to driver pickup (actual wait time) for the Boston experiment. In addition, we report the means of these three variables by race.

Table 3: Balance Test: Relationship between race and features of the Boston trips

Variable	African Americans	White	Difference
Estimated Wait Time	187.805	184.076	-3.728
	(98.895)	(95.785)	(6.755)
Surge Multiplier	1.059	1.058	-0.001
	(0.219)	(0.207)	(0.014)
High AA Density	0.373	0.378	0.005
	(0.484)	(0.485)	(0.032)
High Income	0.397	0.380	-0.018
	(0.490)	(0.486)	(0.032)
Low Pop Density	0.368	0.420	0.052
	(0.483)	(0.494)	(0.033)
Observations	413	498	911

Notes: This table shows the means of the variables when RAs used African American sounding names and white sounding names. Estimated wait time is in seconds.

Table 4: Models of the log of Acceptance Time for UberX and Lyft in the Boston experiment

		UberX			Lyft	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
African American	0.035	0.041	0.023	0.018	0.022	0.013
	(0.081)	(0.081)	(0.081)	(0.091)	(0.091)	(0.091)
$\ln(ETA)$		0.179*	0.199**		0.102	0.116
		(0.093)	(0.094)		(0.091)	(0.092)
Low Pop Density			0.161*			0.104
			(0.084)			(0.094)
High Income			-0.058			0.058
			(0.083)			(0.094)
High AA Density			0.047			-0.090
			(0.083)			(0.093)
Individual Fixed Effects?	No	No	Yes	No	No	Yes
Observations	372	372	372	387	387	387
R-squared	0.001	0.010	0.050	0.000	0.003	0.051

Notes: This table shows regression models of the log of the time it takes until a trip is accepted for the Boston experiment. We split the regressions by service (UberX and Lyft). The first model is a simple difference in mean times across African Americans and white. The second model controls for the estimated wait time prior to the trip being requested. Model 3 includes route characteristics and research assistant fixed effects. For route characteristics, we use demographics from the Census Block Group of the pickup point using 2010 data. "High AA Pop" is defined as 20% or greater African American population, "High Income" is defined as median household income greater than $60k annually, and "Low Population Density" is defined as fewer than 12,800 people per square mile.

Table 5: Models of the log of actual waiting time for UberX and Lyft in the Boston experiment

		UberX			Lyft	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
African American	-0.078	-0.062	-0.053	-0.030	-0.009	-0.005
	(0.065)	(0.061)	(0.061)	(0.067)	(0.061)	(0.062)
$\ln(ETA)$		0.523***	0.538***		0.565***	0.565***
		(0.070)	(0.070)		(0.061)	(0.063)
Low Pop Density			0.148**			0.029
			(0.063)			(0.064)
High Income			-0.012			0.053
			(0.062)			(0.064)
High AA Density			-0.019			-0.095
			(0.062)			(0.063)
Individual Fixed Effects?	No	No	Yes	No	No	Yes
Observations	372	372	372	387	387	387
R-squared	0.004	0.135	0.171	0.001	0.185	0.197

Notes: This table shows regression models of the log of actual wait time from when the trip was requested to the time the research assistant was picked up in the Boston experiment. We split the regressions by service (UberX and Lyft). The first model is a simple difference in mean times across African Americans and white. The second model controls for the estimated wait time prior to the trip being requested. Model 3 includes route characteristics and research assistant fixed effects. For route characteristics, we use demographics from the Census Block Group of the pickup point using 2010 data. "High AA Pop" is defined as 20% or greater African American population, "High Income" is defined as median household income greater than \$60k annually, and "Low Population Density" is defined as fewer than 12,800 people per square mile.

Table 6: Comparisons of cancellation rates between African American and white passengers in the Boston experiment

Passengers	Uber		Lyft	
	African American	White	African American	White
All	10.1%	4.9%	6.0%	7.7%
	(1.9%)	(1.5%)	(1.5%)	(1.9%)
	247	204	251	209
Males	11.2%	4.5%	8.4%	8.7%
	(2.6%)	(2.0%)	(2.3%)	(2.8%)
	152	112	154	115
Females	8.4%	5.4%	2.1%	6.4%
	(2.9%)	(2.4%)	(1.5%)	(2.6%)
	95	92	97	94

Notes: This table reports the mean cancellation rates across the two services, as well as the cancellation rates across the four gender·race pairs in the Boston experiment. The standard error is in parentheses, while the number of observations for each group is below the standard error. The total number of observations is 911 while the total number of cancellations is 66.

Table 7: Cancellation rates were higher for African Americans using Uber, but not Lyft, in Boston

VARIABLES	Lyft Means	Lyft Weekday Effects	Uber Means	Uber Weekday Effects	Uber Rider Effects	Uber Transit Effects	Uber Demographic Interaction	Uber Surge Interaction
African American	-0.017		0.052**					
	(0.024)		(0.025)					
African American Male		0.002		0.073**	0.071**	0.071**	0.003	-0.226*
		(0.033)		(0.034)	(0.035)	(0.035)	(0.042)	(0.137)
White Female		-0.028		0.013	-0.059	-0.031	-0.066	-0.071
		(0.037)		(0.039)	(0.137)	(0.137)	(0.136)	(0.136)
African American Female		-0.057		0.038	-0.042	-0.020	-0.043	-0.049
		(0.037)		(0.039)	(0.143)	(0.142)	(0.142)	(0.142)
High AA Density		-0.013		-0.008	-0.008	-0.008	-0.011	-0.012
		(0.025)		(0.026)	(0.026)	(0.026)	(0.026)	(0.026)
High Income		-0.005		-0.065**	-0.066**	-0.060**	-0.063**	-0.066**
		(0.025)		(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Low Pop Density		-0.009		0.054**	0.055**	0.053**	-0.001	0.001
		(0.025)		(0.026)	(0.026)	(0.026)	(0.032)	(0.032)
Subway						0.102***		
						(0.038)		
AA Male*Low Pop Dens							0.160***	0.163***
							(0.054)	(0.054)
AA Male*Surge Multiplier								0.215*
								(0.123)
Surge Multiplier								-0.046
								(0.057)
Weekday Fixed Effects	No	Yes No	Yes	Yes	Yes	Yes	Yes	Yes
RA Fixed Effects	No	No	No	No	Yes	Yes	Yes	Yes
Observations	460	460	451	451	451	451	451	450
R-squared	0.001	0.015	0.009	0.041	0.050	0.066	0.070	0.076

Notes: This table shows linear probability models for cancellations in the Boston experiment. We first show the results for Lyft requests only. Lyft provides the name and photo of the prospective rider prior to the Lyft driver accepting the ride. Therefore, we would not expect to see an impact of name on the probability of a cancellation since any discrimination that could occur would happen prior to acceptance. An UberX driver does not see the name of the prospective passenger until after accepting the ride. One channel of discrimination, therefore, would be for the driver to accept the ride, receive a signal about the rider's race, and then cancel. The first model is a simple difference in the average rate of cancellations across implied race. We then add day-of-week and pickup location characteristics as controls and allow for gender differences. The third model for Uber adds research assistant fixed effects. The fourth model adds an indicator variable for whether the pickup location is a public transit location in case the attractiveness of these rides is different. The final model interacts the African American male name with an indicator variable for whether the location is in a low-population density area. For route characteristics, we use demographics from the Census Block Group of the pickup point using 2010 data. "High AA Pop" is defined as 20% or greater African American population, "High Income" is defined as median household income greater than $60k annually, and "Low Population Density" is defined as fewer than 12,800 people per square mile. The surge multiplier ranges from 1 to 2.6. For trips that were cancelled, RAs recorded the surge multiplier of the next requested trip. We use this multiplier for cancelled trip observations. We are unable to reject equality between the female African American coefficients and either the male African American or the female white coefficients.

Table 8: Effects on waiting times for African American males in low-density areas.

VARIABLES	(1) Uber Acceptance Time	(2) Uber Wait Time
African American	-0.046	-0.085
	(0.088)	(0.066)
AA Male*Low Pop Dens	0.312**	0.144
	(0.158)	(0.119)
High AA Density	0.045	-0.020
	(0.083)	(0.062)
High Income	-0.059	-0.012
	(0.083)	(0.062)
Low Pop Density	0.055	0.100
	(0.099)	(0.074)
Observations	372	372
R-squared	0.060	0.175

Table 9: Test of race effect on (log) acceptance time and (log) waiting time in the Seattle pilot

Service	Metric	Estimated African American Effect	p value
Uber	Acceptance Time	0.24	0.23
	Actual waiting time	0.28	0.17
Lyft	Acceptance Time	0.17	0.20
	Actual waiting time	-0.04	0.63
Flywheel	Acceptance Time	0.27	0.17
	Actual waiting time	0.05	0.65

The estimated African American effects are calculated based on the linear regression models after controlling for the estimated waiting time, route fixed effects, and pickup location neighborhood income, population density, and the proportion of African American population. The p-values are based on a randomization test in Young (2018), where we run a Monte Carlo and for each iteration, RAs are randomly assigned race categories and the regression is re-estimated. The p-values are calculated as the percentage of time the African American coefficient is larger than the non-RA randomized coefficient.

References

ANGRIST, J. D., S. CALDWELL, AND J. V. HALL (2017): "Uber vs. Taxi: A Driver's Eye View," Working Paper 23891, National Bureau of Economic Research.

AYRES, I., M. BANAJI, AND C. JOLLS (2015): "Race effects on eBay," *The RAND Journal of Economics*, 46, 891–917.

BERTRAND, M. AND S. MULLAINATHAN (2004): "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *The American Economic Review*, 94, 991–1013.

BROWN, A. E. (2018): "Ridehail revolution: Ridehail travel and equity in Los Angeles," Ph.D. thesis, UCLA.

CHEN, M. K., J. A. CHEVALIER, P. E. ROSSI, AND E. OEHLSEN (2017): "The value of flexible work: Evidence from uber drivers," Tech. rep., National Bureau of Economic Research.

COHEN, P., R. HAHN, J. HALL, S. LEVITT, AND R. METCALFE (2016): "Using big data to estimate consumer surplus: The case of uber," Tech. rep., National Bureau of Economic Research.

CRAMER, J. AND A. B. KRUEGER (2016): "Disruptive change in the taxi business: The case of Uber," *American Economic Review*, 106, 177–82.

DOLEAC, J. L. AND L. C. STEIN (2013): "The Visible Hand: Race and Online Market Outcomes," *The Economic Journal*, 123, F469–F492.

DONNELLY, A. (2015): "Aboriginal actor Jack Charles 'refused taxi twice in three days'," *BBC News*.

EDELMAN, B. AND M. LUCA (2014): "Digital Discrimination: The Case of Airbnb.com," Tech. rep., Harvard University.

EDELMAN, B., M. LUCA, AND D. SVIRSKY (2016): "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment," Tech. rep., Harvard University.

FRYER, R. G. AND S. D. LEVITT (2004): "The Causes and Consequences of Distinctively Black Names," *The Quarterly Journal of Economics*, 119, 767–805.

GE, Y., C. R. KNITTEL, D. MACKENZIE, AND S. ZOEPF (2016): "Racial and Gender Discrimination in Transportation Network Companies," Working Paper 22776, National Bureau of Economic Research.

GLANVILLE, D. (2015): "Why I Still Get Shunned by Taxi Drivers," *The Atlantic*, October 24.

GODDARD, T., K. B. KAHN, AND A. ADKINS (2015): "Racial bias in driver yielding behavior at crosswalks," *Transportation Research Part F: Traffic Psychology and Behaviour*, 33, 1–6.

GONEN, Y. (2015): "Puerto Rican mayor booted from NYC cab | New York Post," *NY Daily News*.

HALL, J. V. AND A. B. KRUEGER (2018): "An analysis of the labor market for Uber's driver-partners in the United States," *ILR Review*, 71, 705–732.

HARSHBARGER, R. (2015): "City puts biased drivers who refuse rides on notice with new video," *New York Post*, June 9th.

HUGHES, R. AND D. MACKENZIE (2016): "Transportation network company wait times in Greater Seattle, and relationship to socioeconomic indicators," *Journal of Transport Geography*, 56, 36–44.

NICHOLSON, E. (2013): "Southern Dallas Leaders Say Uber is Profiling Customers, and They Want City Hall to Act," *Dallas Observer*, October 21.

PIFF, P. K., S. DM, C. S, M.-D. R, AND K. D (2012): "Higher Social Class Predicts Increased Unethical Behavior," *Proceedings of the National Academy of Science*, 109, 4086–4091.

POPE, D. G. AND J. R. SYDNOR (2011): "What's in a Picture? Evidence of Discrimination from Prosper. com," *Journal of Human Resources*, 46, 53–92.

SMART, R., B. ROWE, A. HAWKEN, M. KLEIMAN, N. MLADENOVIC, P. GEHRED, AND C. MANNING (2015): "Faster and Cheaper: How Ride-Sourcing Fills a Gap in Low-Income Los Angeles Neighborhoods," Tech. rep., BOTEC Analysis Corporation.

UBERPEOPLE.NET (2015): "Uber Cancellation Abuse - Mini FAQ," *Uber Drivers Forum*.

YOUNG, A. (2018): "Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results," *The Quarterly Journal of Economics*, 134, 557–598.

A Online Appendix

Figure A.1: Routes, pick-up locations and neighborhood characteristics of Seattle experiment.

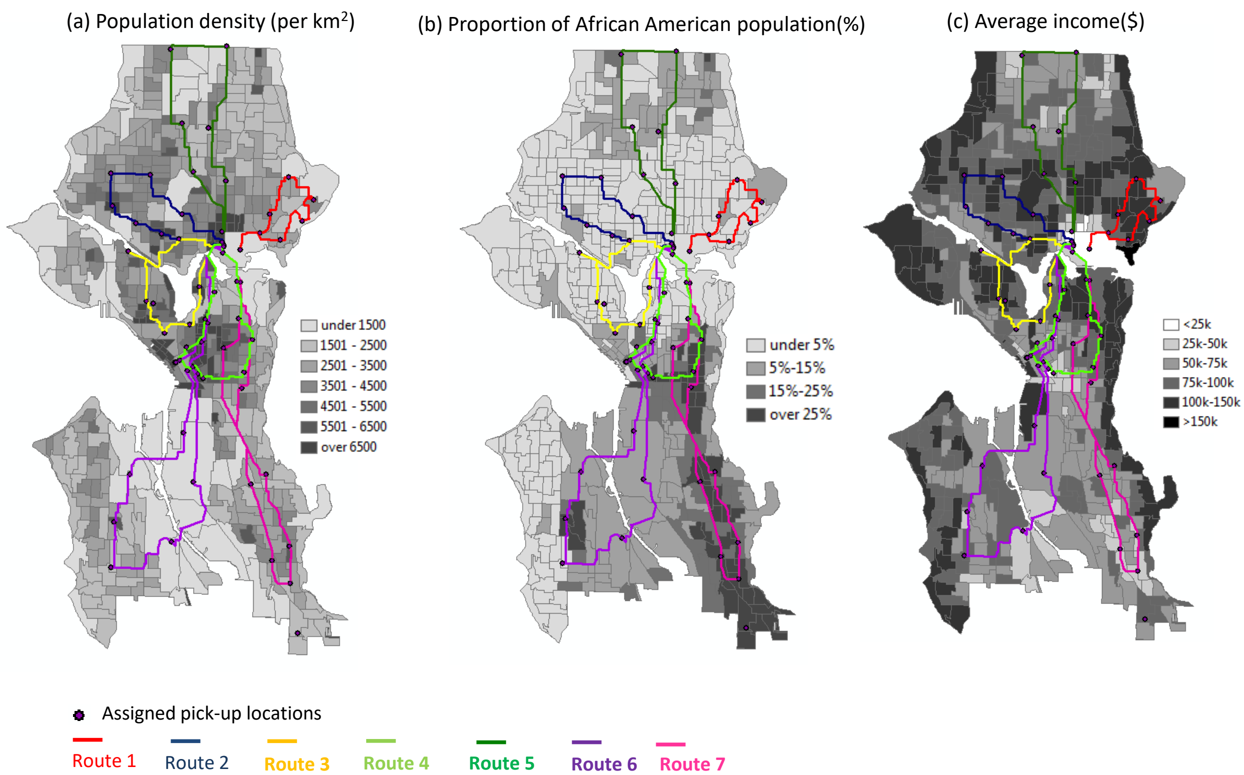


Table A.1: Attribute variables of experiment design

Attributes	**Attribute levels**
Routes	7 levels: 7 routes
Mode for first trip	3/2 levels: UberX, Lyft, & Flywheel in stage 1; UberX & Lyft in stage 2
Race	2 levels: African American, white
Sex	2 levels: female, male
Day of the week	4 levels: Monday, Tuesday, Wednesday, Thursday
Week of data collection	4/2 levels: 4 weeks in stage 1, 2 weeks in stage 2

Notes: This table shows the number of routes table by research assistants, the mode selection, research assistant race and gender, and the frequency of trips taken for the Seattle pilot.

Figure A.2: Data collection process illustration.

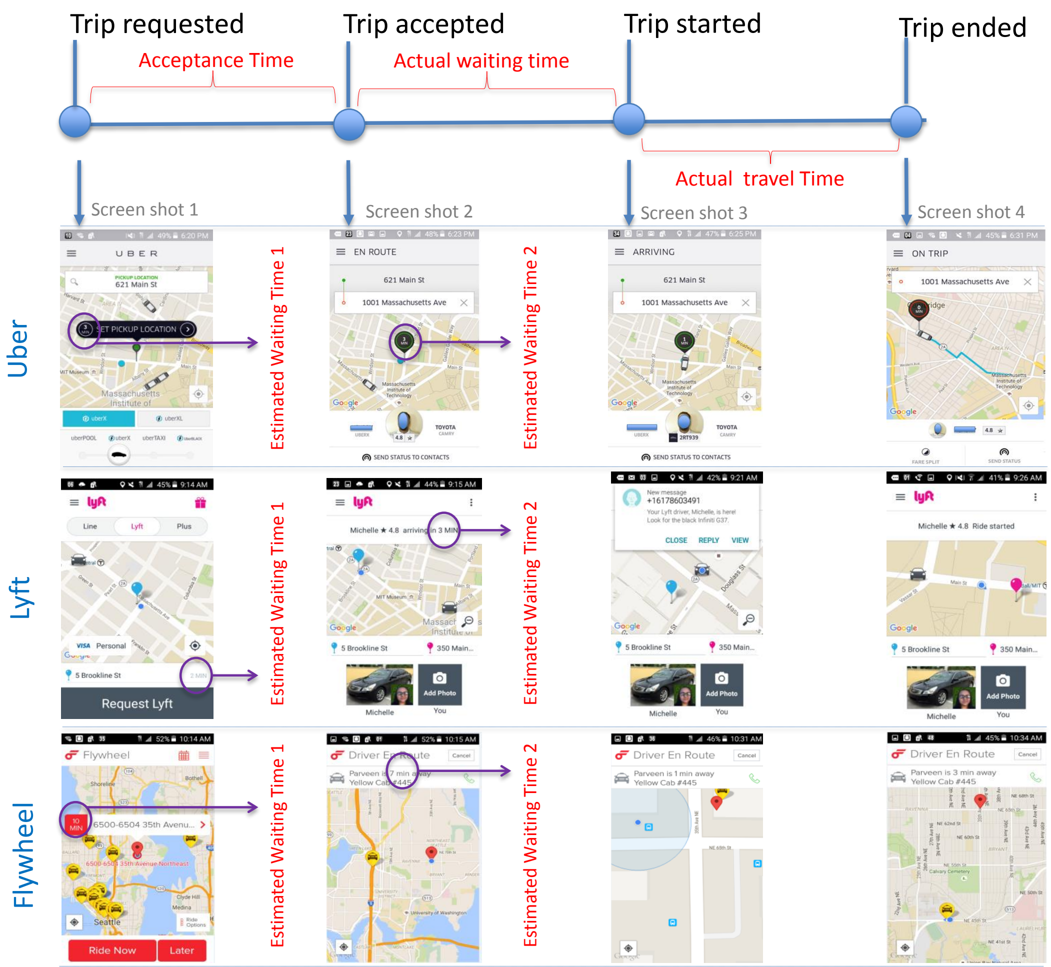
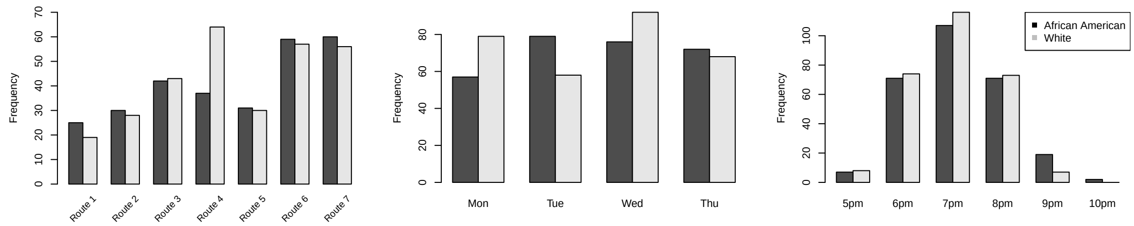


Figure A.3: Balance in days, times, and tour IDs for trips taken by African American and white RAs in Seattle.



Notes: One African American male RA dropped out after four weeks of data collection for the Seattle experiment, which is why we have less Route 4 trips by African American RAs than white RAs. To address the concerns of this imbalance problem, we checked the robustness of the analyses shown in Table 4 and Table 5 by comparing the results with models based on the following adjusted datasets: (1) we deleted all Route 4 trips; (2) We randomly deleted extra trip tours on Route 4 by white RAs. The race effects based on these adjusted datasets are generally consistent with the model results.

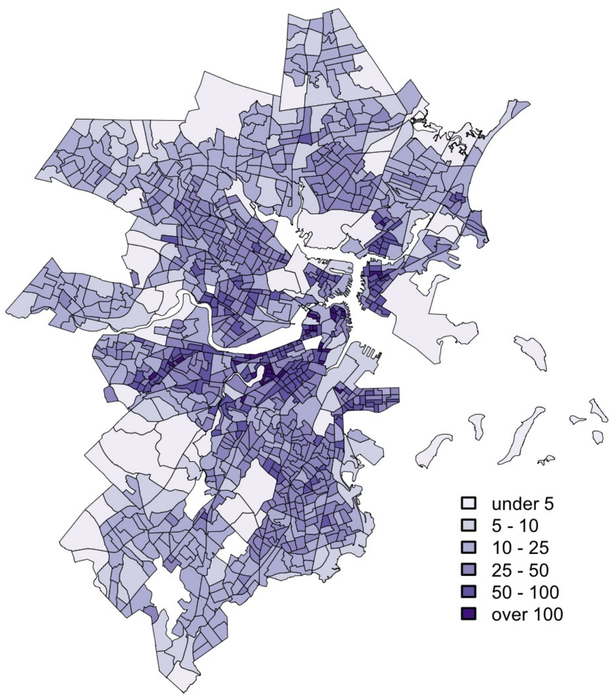
Table A.2: Descriptive analysis of the acceptance time, estimated waiting time 1 and actual waiting time of Uber, Lyft and Flywheel in the Seattle pilot

Service	Attributes	Acceptance time(s)	Est. waiting time 1 (min)	Actual waiting time (s)
Uber	Mean	25	4.1	275
	S.D.	21	1.9	212
	Median	18	4	225
	African American Mean	29	4.3	316
	White Mean	21	3.9	240
	Num.obs	190	200	187
Lyft	Mean	23	3.0	284
	S.D.	24	1.8	156
	Median	16	3	270
	African American Mean	23	3.2	284
	White Mean	19	2.7	284
	Num.obs	199	199	202
Flywheel	Mean	44	6.2	446
	S.D.	69	3.5	263
	Median	28	5	381
	African American Mean	35	6.4	447
	White Mean	35	6.1	444
	Num.obs	130	124	124

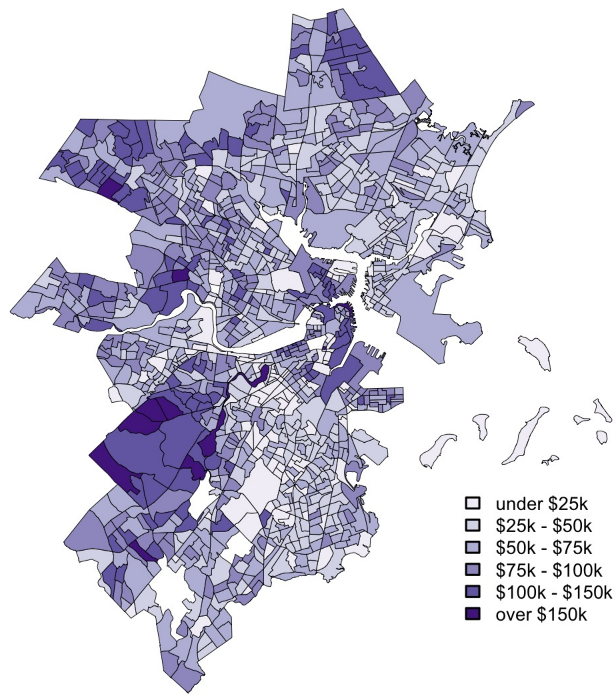
Notes: This table shows the number of trips taken and summary statistics for the time until the request was accepted (Acceptance time), the *estimated* wait time reported by the app when the trip was requested (Est. wait time 1), and the actual wait time from when the trip was requested (Actual wait time) for the Seattle pilot. In addition, we report the means of these three variables by race.

Figure A.4: GIS Analysis of Boston-area Census Block Groups by % African American Residents, Average Household Income, and Population Density

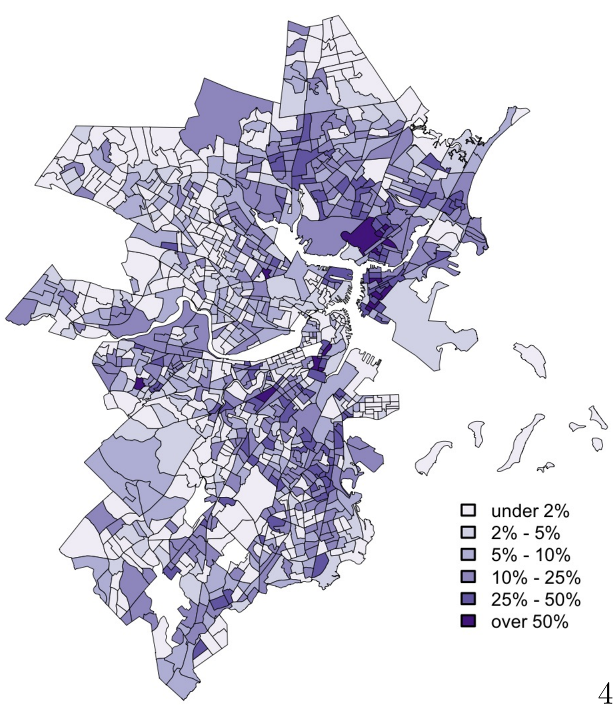
(a) Population Density, Thousands / Square Mile



(b) Median Household Income



(c) Share Non-English Speaking Households



(d) Black Share of Population

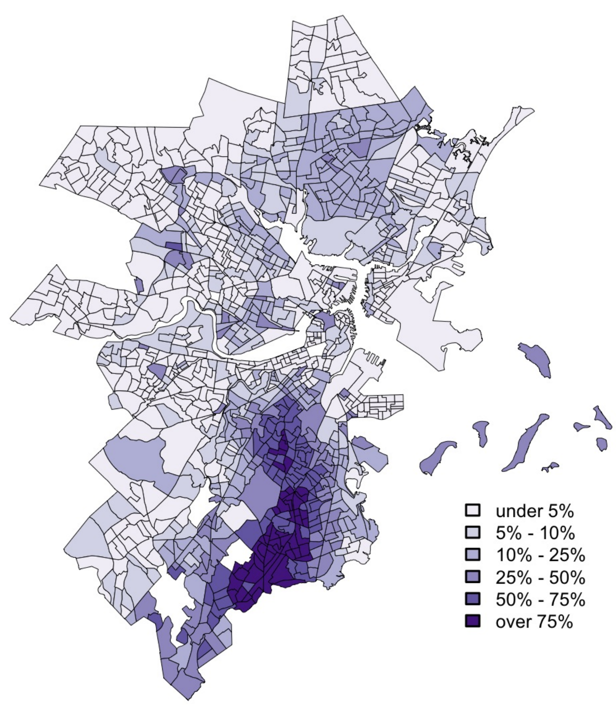


Figure A.5: Quantile regression coefficients of the effect of race on the log of acceptance time

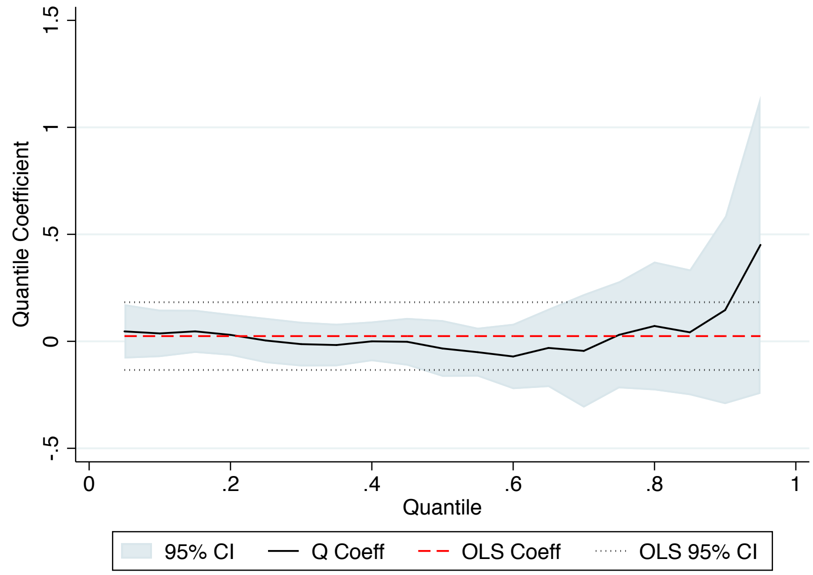


Figure A.6: Quantile regression coefficients of the effect of race on the log of total time

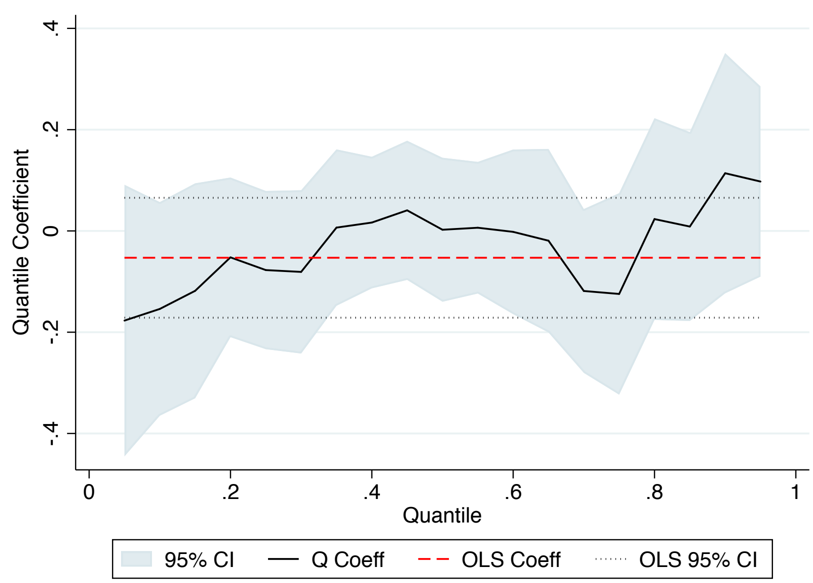


Figure A.7: Quantile regression coefficients of the effect of gender on the log of drive distance

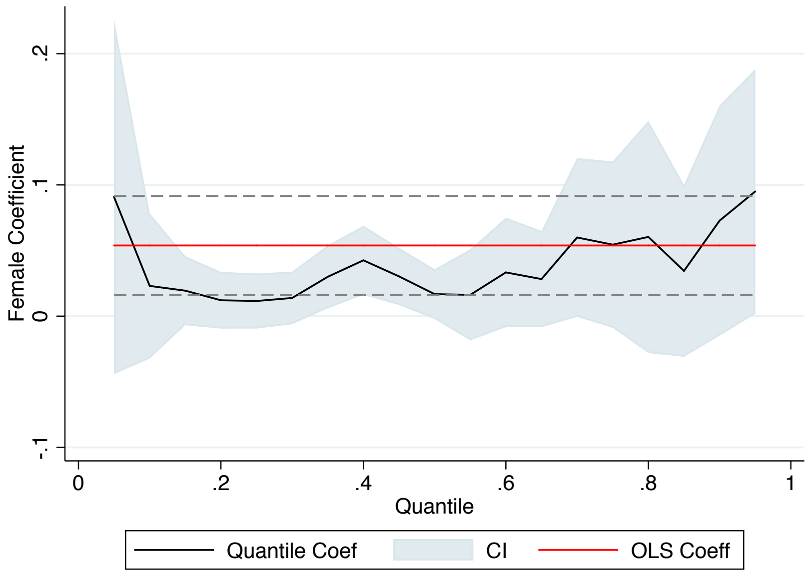


Table A.3: Balance Test: Relationship between race and the log of *estimated* waiting times prior to requesting the trip in the Seattle pilot

	UberX	Lyft	Flywheel
(Intercept)	1.27***	0.85***	1.68***
	(0.04)	(0.06)	(0.08)
African American Traveler	0.08	0.12	−0.04
	(0.06)	(0.09)	(0.11)
R^2	0.01	0.01	0.00
Adj. R^2	0.00	0.01	-0.01
Num. obs.	188	183	121
RMSE	0.43	0.59	0.62

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Notes: This table shows the affect of race on *estimated* wait time prior to the trip request in the Seattle pilot. We find no systematic relationship between the service's estimated wait times and race.

Figure A.8: Simulated African American effects coming from random assignment of Boston RA fixed effects

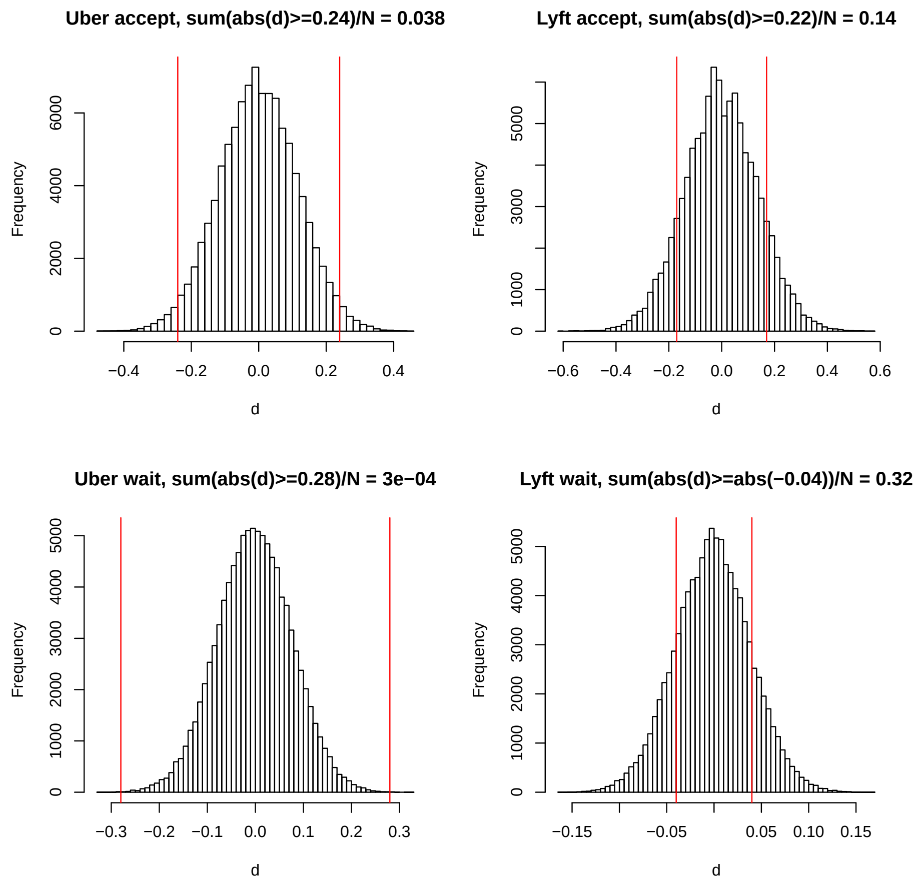


Figure A.9: Randomization Inference Results of African American Effects on Accept time and Waiting time of Seattle Pilot Experiment

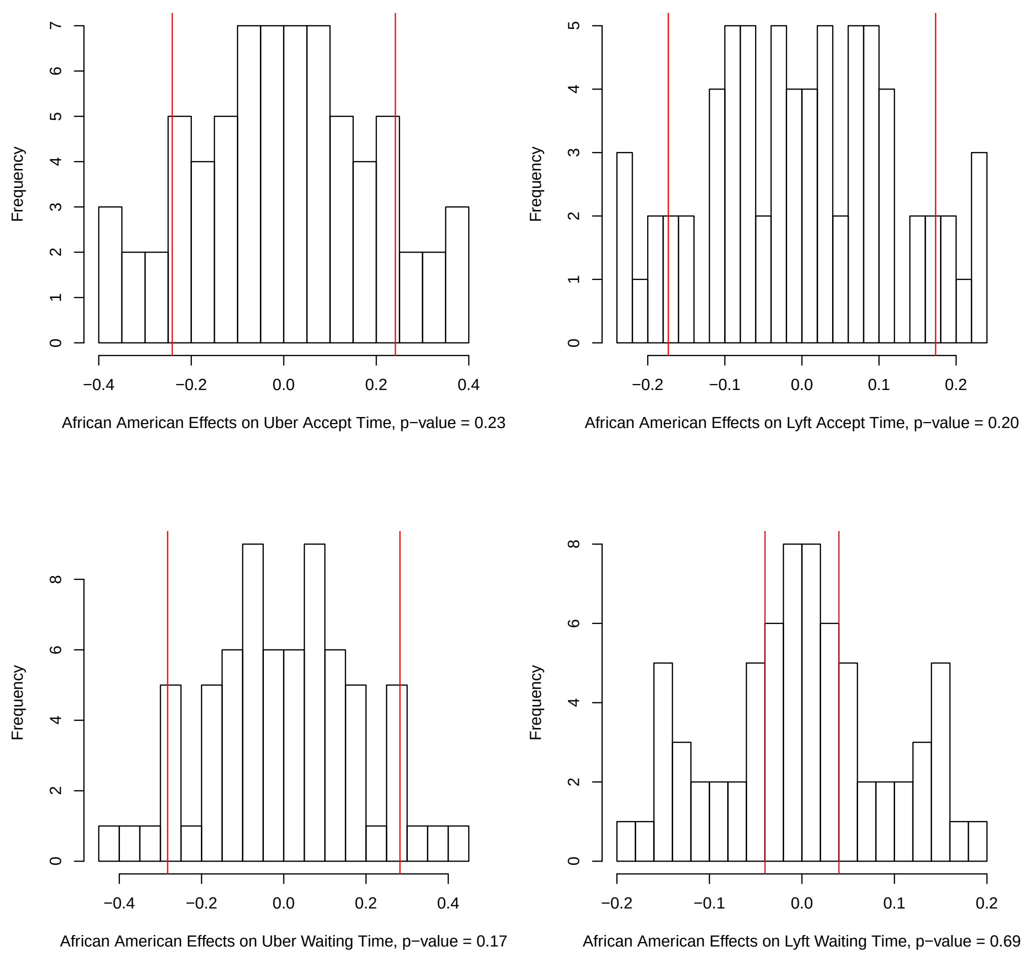


Figure A.10: Number of taxis passing by African American and white RAs while hailing from the start

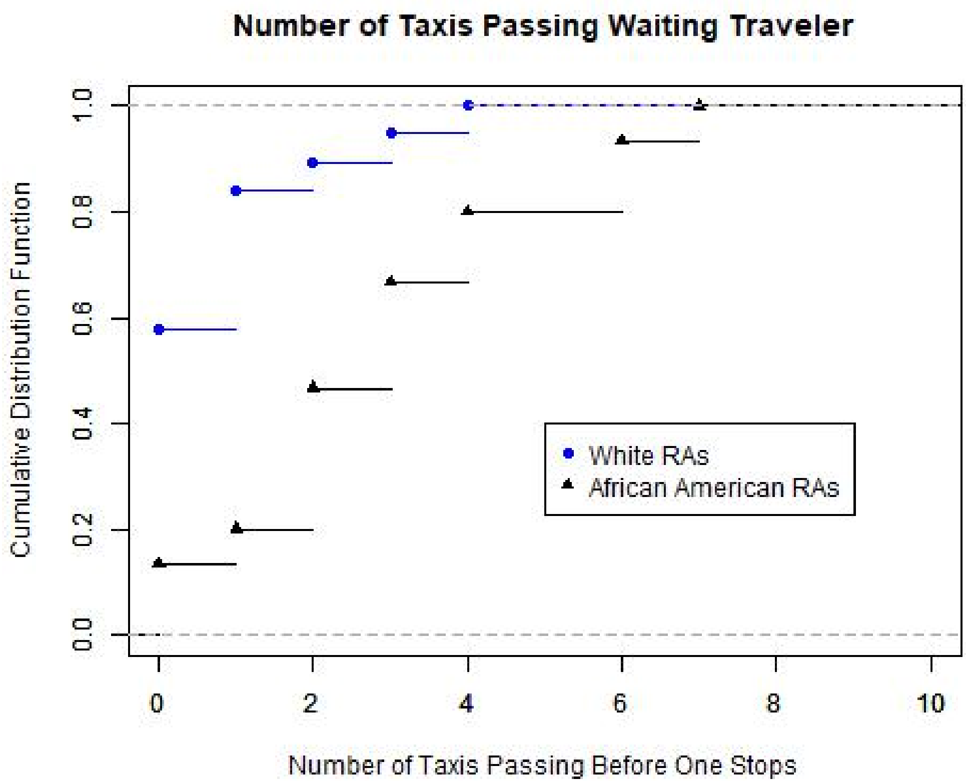


Figure A.11: Cumulative Star Ratings for African American and White RAs Using UberX in Seattle.

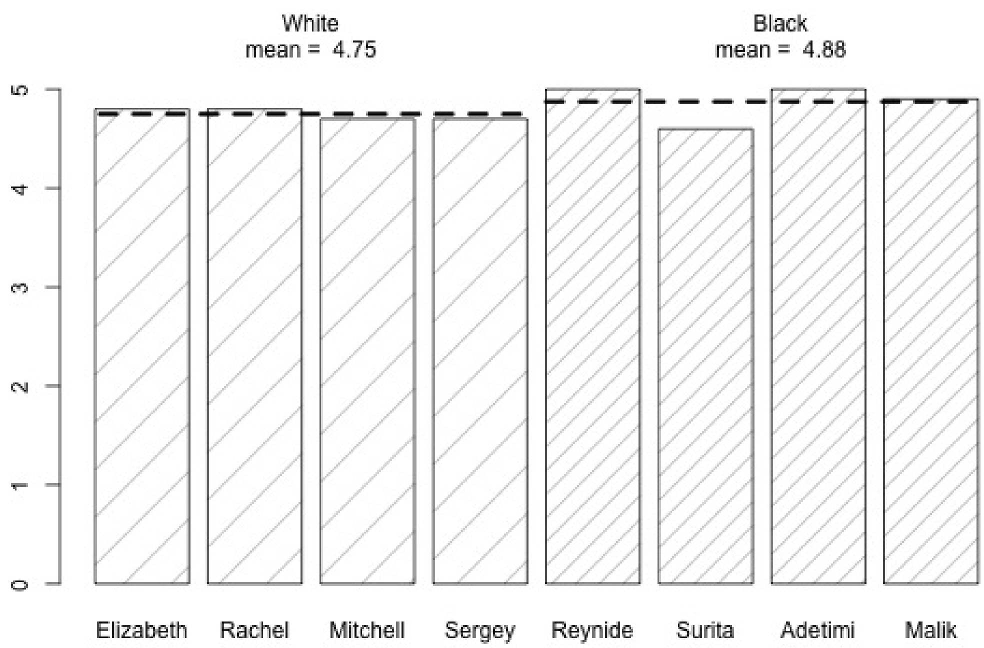


Figure A.12: Cumulative Star Ratings for African American and White RAs Using Lyft in Seattle.

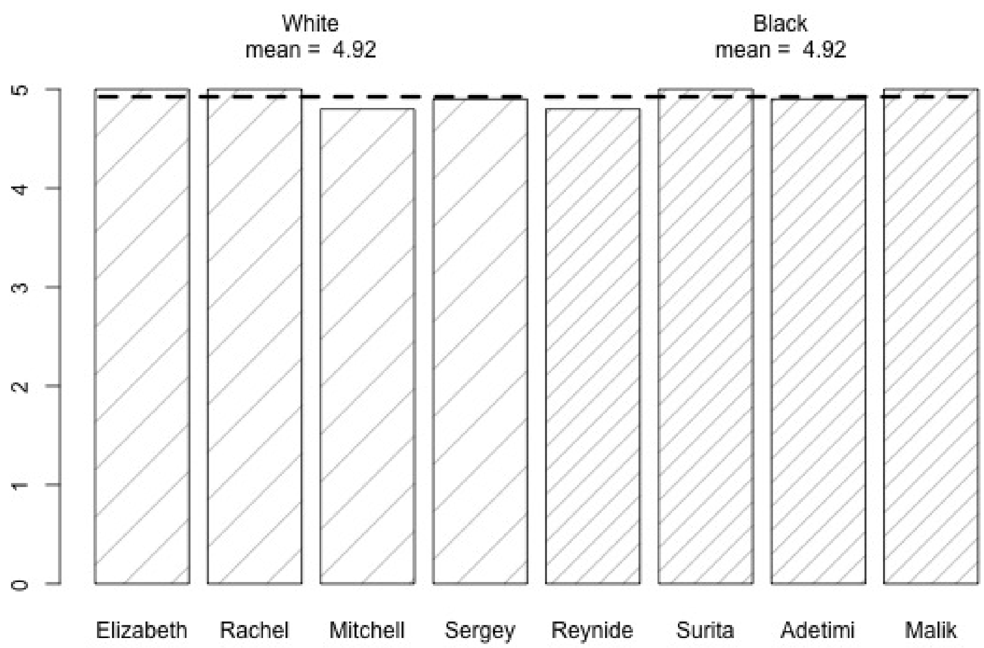


Figure A.13: Cumulative Star Ratings for African American and White RAs Using UberX in Boston.

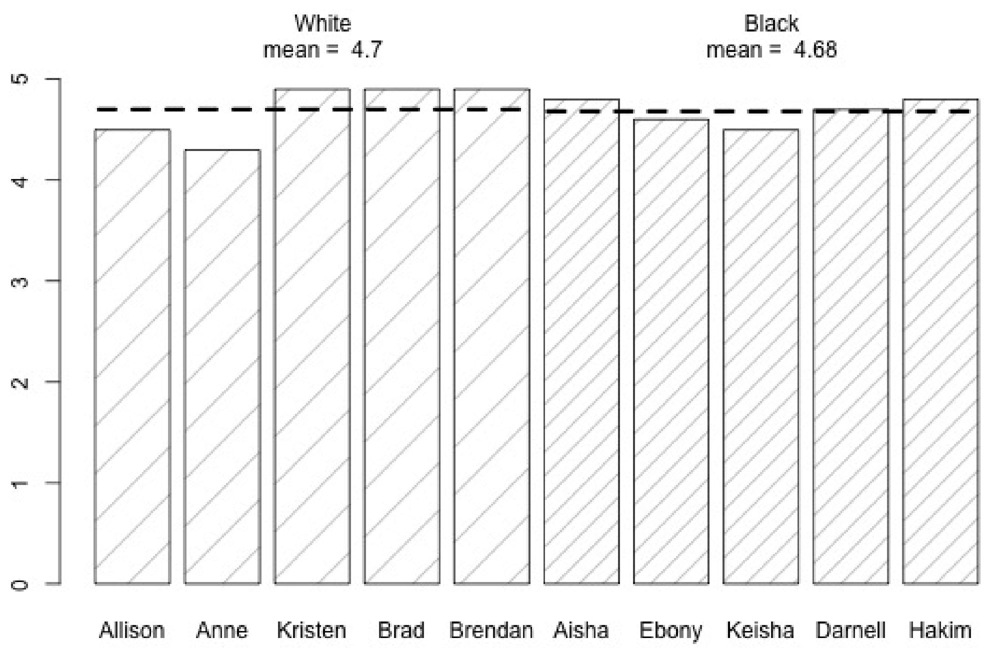


Figure A.14: Cumulative Star Ratings for African American and White RAs Using Lyft in Boston.

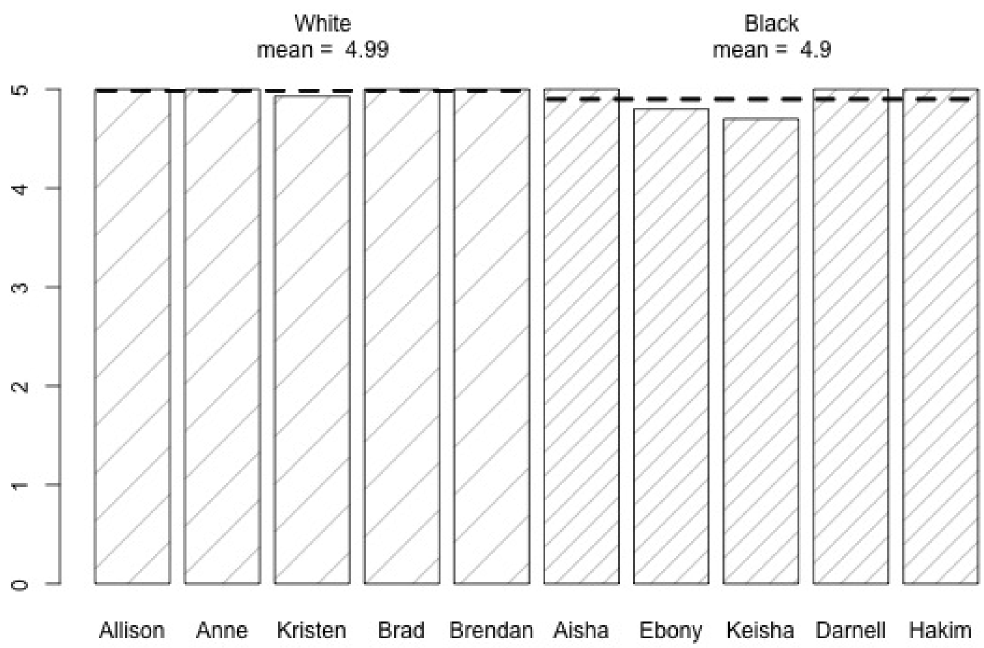


Table A.4: Required sample sizes

Variable	5 percent	10 percent	15 percent
Actual Wait Time	1,736	434	193
Estimated Wait Time	903	226	101
Ratio of Actual To Estimated Wait Time	1,378	345	154
Δ Actual and Estimated Wait Time	37,672	9,418	4,186
Cancellation	NA	NA	NA
Actual Drive Time	1,056	264	118
Estimated Drive Time	NA	NA	NA
Ratio of Actual to Estimated Drive Time	NA	NA	NA
Δ Actual and Estimated Drive Time	NA	NA	NA