Online Appendix for "When Transparency Fails: Bias and Financial Incentives in Ridesharing Platforms"

Jorge Mejia
Kelley School of Business, Indiana University, Bloomington, IN 47405, jmmejia@iu.edu, http://go.iu.edu/jorge

Chris Parker
Kogod School of Business, American University, Washington, DC 20016, chris.parker@american.edu, https://www.american.edu/kogod/faculty/christop.cfm

1. Introduction

In this online appendix we describe a previous version of the experiment. There are three main differences between the experiments:

1. **LGBT Treatment** In this version we ran two experiments back-to-back. In the first, all rides were requested without the LGBT support treatment. All rides in the second experiment had the LGBT support treatment. In this regard we could not disentangle timing of the experiments from the actual LGBT support treatment.
2. **Timing** In this version we ran the experiments in October/November 2017. The main version ran in October/November 2018.
3. **Names** In this version we use only one name per treatment: Keisha (URM Female), Rasheed (URM Male), Emily (Caucasian Female), and Brad (Caucasian Male). The main version of the experiment used these names but added an additional name to each treatment group to ensure that the name was not driving the observed results.
4. **Analysis** Analysis for this version used simple t-tests for comparison of means. The main experiment utilizes proportions tests and logistic regression to evaluate differences.

2. Experimental Design

Below we describe the previous experiment performed to test whether the operational transparency change eliminated bias and the effects of dynamic pricing on that bias. We also detail the three measures of bias we use.

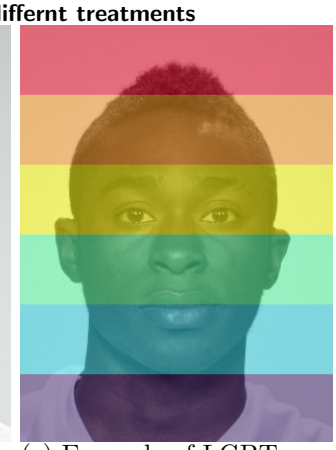
2.1. Treatment Groups

Our experiment contains four treatments each with two possible scenarios for a total of sixteen distinct treatment cells. Three of the treatments alter rider characteristics. Two of these treatments vary names and profile pictures to show that the rider is likely to be 1) Female or Male (F/M) and 2) Caucasian or URM (Cau/URM). The third treatment uses a rainbow-colored overlay on

Figure 1 Images and names used for the differnt treatments

(a) Keisha (URM Female)

(b) Rasheed (URM Male)



(c) Example of LGBT Support.

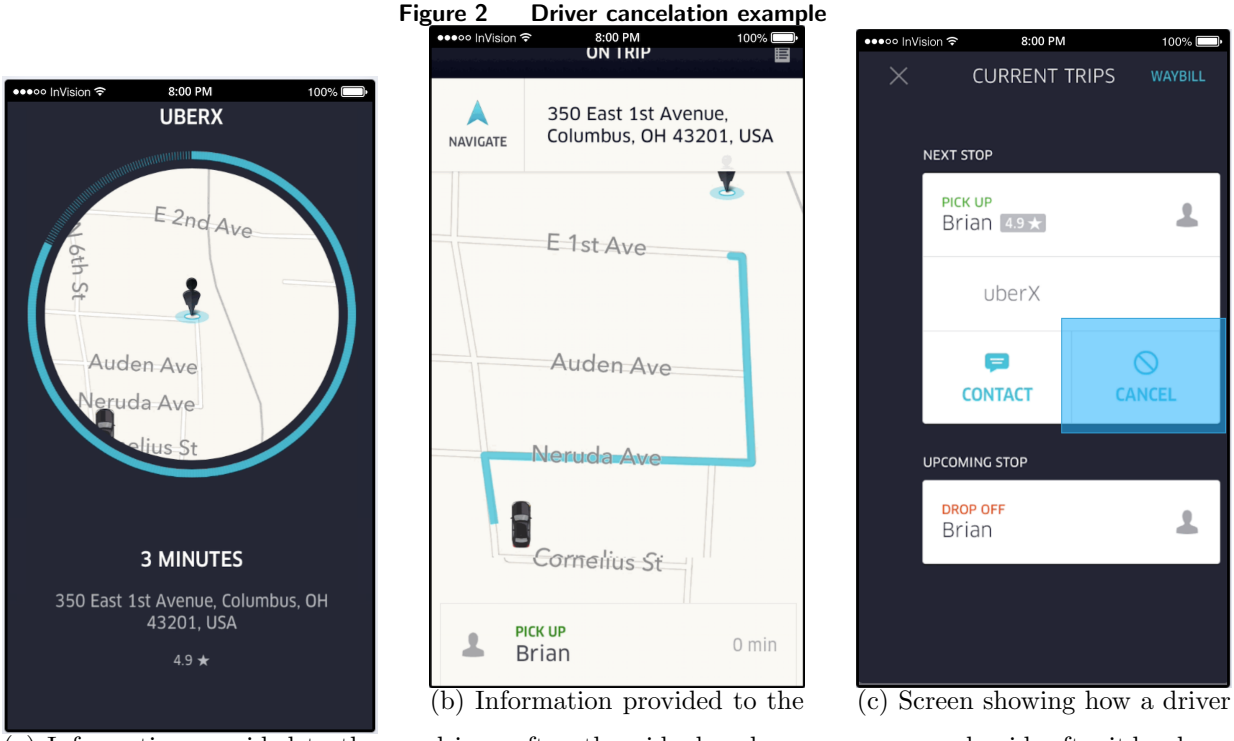
(d) Emily (Caucasian Female)

(e) Brad (Caucasian Male)

the profile picture to signal that the rider is an ally of the LGBT community (LGBT/Non-LGBT). Figure 1 shows the names and pictures used for the different ethnicity and gender treatments as well as an example of the rainbow filter applied for the LGBT support treatment. At least 92% of people perceive Emily and Brad as Caucasian names and Keisha and Rasheed to be black names (Bertrand and Mullainathan 2004). No last names were used as this information is never available to the driver. The pictures come from The Face Database (Minear and Park 2004). The final treatment is used to examine whether pricing amplifies or alleviates biases on ridesharing platforms. In this treatment, we vary the time of day to be at a time when there is not typically much demand such that prices are likely to be low or at a time when there is high demand such that prices are likely to be high (Non-Peak/Peak).

The experiment is run using a major ridesharing platform in Washington, DC in October/November 2017. The pickup is fixed at a central Metro stop, the drop-off is fixed at an airport, and the rider's rating is 4.8 throughout.

Figure 2 Driver cancelation example



(a) Information provided to the driver at the ride request stage.

(b) Information provided to the driver after the ride has been accepted.

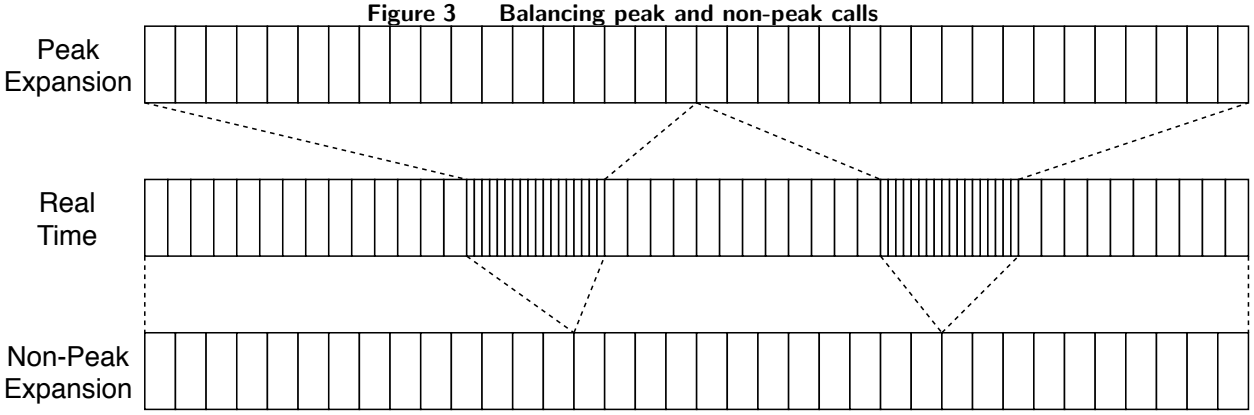
(c) Screen showing how a driver can cancel a ride after it has been accepted.

Note. Example of what an Uber driver would see during the request process (a), after accepting a ride (b), and when canceling a ride (c).

The experiment was performed over two phases. In the first phase we altered gender, ethnicity, and timing of ride requests ignoring LGBT support (Non-LGBT). We did this to be consistent with previous research. The second phase was identical to the first except that all requests included the rainbow filter (LGBT) treatment. Within each phase, a treatment cell is randomly selected by the researchers' computer program from the possibilities described above. Then a ride request is generated through the platform's API with the manipulated rider characteristics corresponding to the treatment cell. The platforms offers a driver (i.e., the subject in this study) the ride at which point the driver decides to accept or reject the ride. As the experiment is implemented after the platform change, the driver can only observe the general location of the rider at this stage. If the driver rejects the ride, then the platform offers the ride to a new driver. This continues until a driver accepts the ride at which point the app notifies the API that the driver is on the way.

Once the ride is confirmed, the rider's name and profile picture become visible to the driver as well as the exact pickup location. We wait for three minutes to allow the driver to cancel if they no longer want to accept the rider. If the driver cancels the ride, they receive no compensation. If the driver has not canceled after three minutes, we cancel the ride, and the driver receives

Figure 3 Balancing peak and non-peak calls



Note. Non-peak and peak calls are spaced at different intervals to account for the fact that peak times are a smaller part of the day than non-peak times.

compensation in the form of a cancelation fee, which ranges from $5 to $10. They are paid at the end of the interaction through the platform as with any other ride and in line with the cancelation policies established for both riders and drivers on the platform. Figure 2 shows an example of the rider information available to Uber drivers at different points in time and how those drivers cancel a ride. The platform on which the experiment was performed has a similar interface.

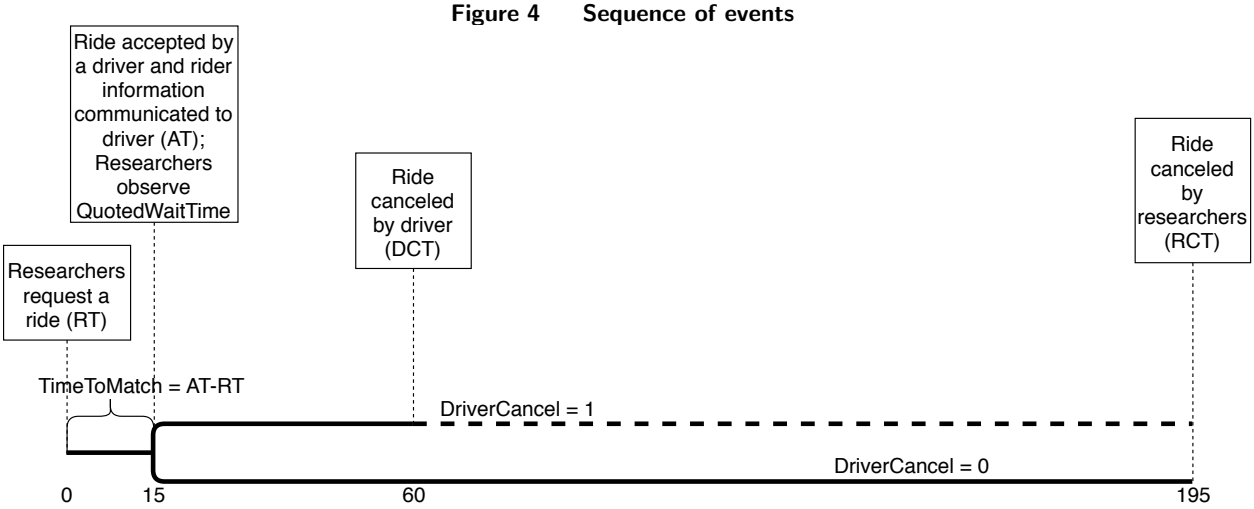
We request 800 rides during Peak and Non-Peak times across our experiment (for a total of 1,600 observations). However, there are not an equal number of peak and non-peak hours each day; AM Peak is 7-10am and PM Peak is 4-7pm for a total of six peak and eighteen non-peak hours each day. During peak hours we call once every ten minutes for a total of 36 peak observations per day. In order to get 36 non-peak observations per day, we need to make 36 ride requests during the 18 non-peak hours, which means we request a ride every 30 minutes.[1] Spreading out the calls in this way has the added benefit of reducing the likelihood that a driver is called two times in a row. Institutional Review Board (IRB) restrictions prevent us from recording any information about the drivers so we cannot explicitly measure how often this occurs.

Figure 3 shows how ride requests are timed throughout the day. The middle row shows actual time, where midnight is both the left and the right side of the row. A call is made in each of the 72 blocks in the day. During the AM and PM peaks, rides are requested at a faster rate as indicated by the skinnier blocks. However, when we zoom into both the peak (top row) and non-peak (bottom row) hours, both have 36 blocks each that are equally spread out across time. Rides were requested for a period of just over 22 days during the fall of 2017 to ensure that there were 100 ride requests placed for each of the sixteen cells for a total of 1,600 observations.[2]

[1] Technically, we call a ride randomly in minutes 1-7 of each ten minute period during peak and in minutes 1-27 during non-peak times. The additional three minutes allows us to cancel the ride before the next 10 minute or 30 minute period begins.

[2] We requested 1,600 rides. However, a small number of requests were not matched to a driver within three minutes of our ride request and therefore were canceled.

Figure 4 Sequence of events



Note. Sequence of making a ride request. The top line is when the driver cancels the ride and the bottom line is when the researchers cancel the ride.

2.2. Measures of Bias

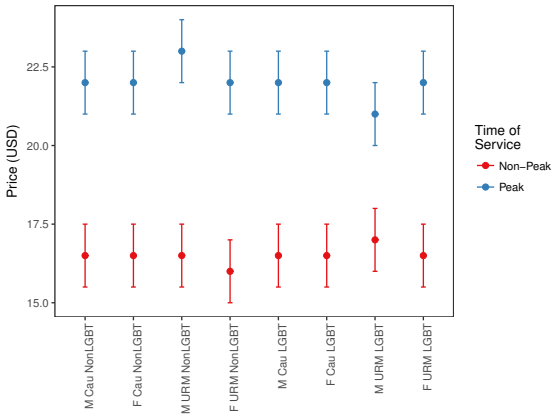
Bias in ridesharing can be exhibited in several ways (Ge et al. 2016). Our study is particularly concerned with drivers either (1) refusing to accept ride requests from certain types of riders, or (2) canceling a ride once rider characteristics are shared.[3] We measure these biased behaviors through three operationalizations of the quality of the service provided, each of which is connected to one of the biased behaviors above: 1a) the time between when a ride is requested and when a ride is confirmed $TimeToMatch = AcceptTime - RequestTime$, 1b) the initial expected waiting time quoted to the rider upon receiving a confirmed ride $QuotedWaitTime$, and 2) whether or not the driver canceled the ride $DriverCancel$, a dummy variable taking the value of 1 if a driver canceled within three minutes of accepting the ride and zero otherwise. Differences in these measures between the different cells are an estimate of the implicit bias towards that group. Figure 4 shows how the sequence of making a ride request occurs and how the metrics above are calculated.

3. Results

Before examining driver-related bias, we first demonstrate that the pricing algorithm does not exhibit any price differences across the different rider-related treatments. Figure 5 shows that quoted prices are similar across the different treatments and are indeed higher during peak times than in the non-peak times.

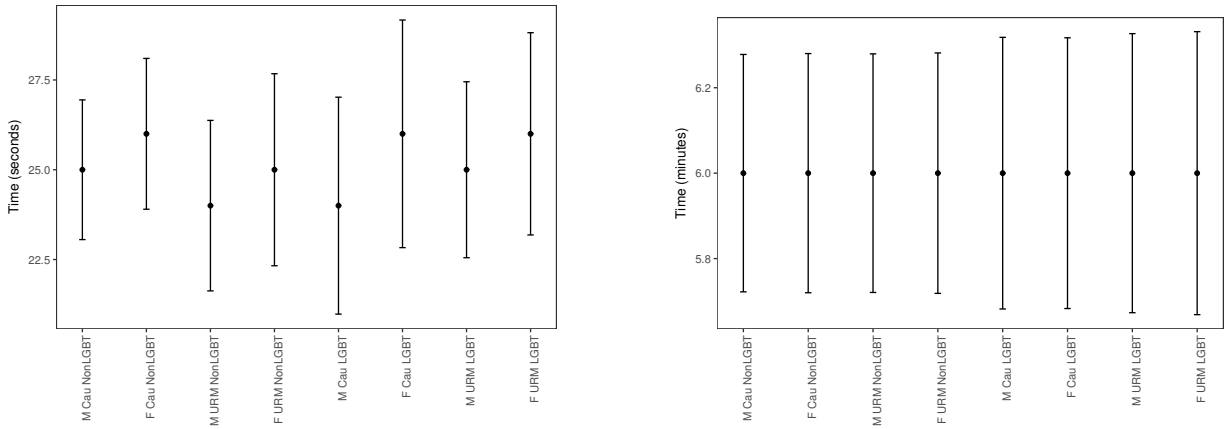
[3] We also measure the time to cancel conditional on the driver canceling. However, there are only a few observations for each of the treatment cells making an analysis of this measure of bias largely uninformative.

Figure 5 Distribution of quoted prices for Peak vs. Non-Peak times



Note. Quoted prices are statistically the same across the different treatments. Bars are the length of one standard deviation due to the data being granular.

Figure 6 Bias metrics at the ride request stage



(a) $TimeToMatch$ (b) $QuotedWaitTime$

Note. Neither $TimeToMatch$ nor $QuotedWaitTime$ exhibit significant differences across treatments. Bars show the 95% confidence interval.

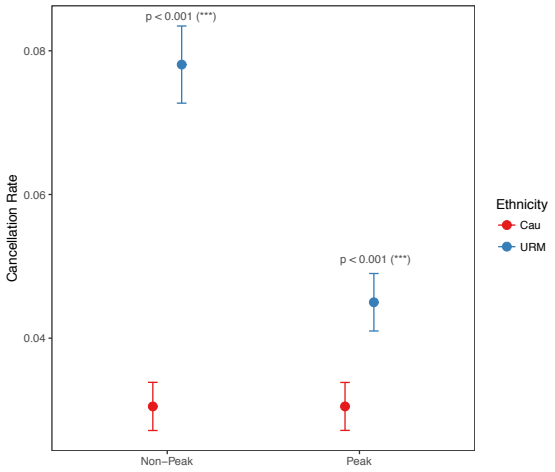
Having shown that there are no price differences, we now turn to examining the drivers' bias. The field experiment, conducted after the platform change, demonstrates that the bias previously observed through higher $TimeToMatch$ and $QuotedWaitTime$ metrics has been completely eliminated. Figure 6a shows that $TimeToMatch$ is close to 25 seconds for all of the treatments while, Figure 6b shows that $QuotedWaitTime$ is close to six minutes for all of the treatments. None of the differences are statistically signifcant.

Table 1 Cancelation Rates Across Treatments

Ethnicity	LGBT Support	Time	Rides Requested	Cancelled	Rate	Std. Dev.
Cau	No	Non-Peak	197	6	0.030	0.024
Cau	Yes	Non-Peak	198	7	0.035	0.026
Cau	No	Peak	198	6	0.030	0.024
Cau	Yes	Peak	195	6	0.031	0.024
URM	No	Non-Peak	192	15	0.078	0.038
URM	Yes	Non-Peak	196	17	0.087	0.039
URM	No	Peak	199	9	0.045	0.029
URM	Yes	Peak	192	9	0.047	0.030

Summary of cancelation rates ignoring the gender dimension.

Figure 7 Driver cancelation rates for different rider ethnicities



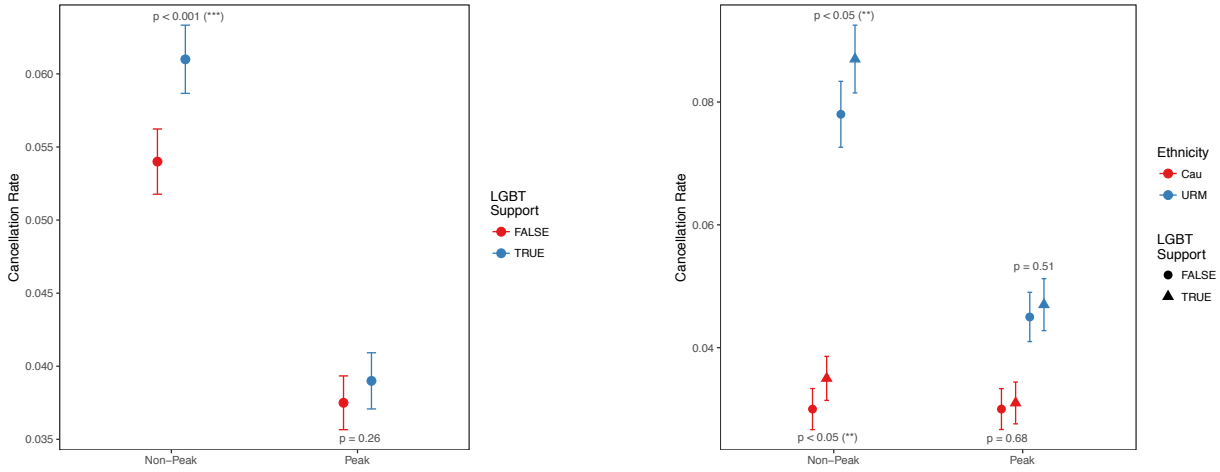
Note. Cancelation rates for riders of different ethnicities in different peak times. Bars show the 95% confidence interval.

While removing bias from the ride request stage is an achievement in and of itself, it does not mean that bias has been removed from the platform. Indeed, we find evidence of bias against URMs and LGBT supporters once a rider is confirmed. Cancelation rates across different treatments (aggregated across Gender) are in Table 1. Below we examine these differences visually and include p-values from a t-test for equal means directly on the figures.

Figure 7 shows the results limited to varying ethnicity and time of service. Cancelation rates are higher for URMs than Caucasians during both Peak and Non-Peak times. Comparing across Peak and Non-Peak treatments, we observe that the pricing effect dominates, i.e., cancelation rates are lower during Peak times for URMs.

The results also demonstrate that there is a bias for supporting the LGBT community as exhibited by higher cancelation rates during Non-Peak times. However, as before this bias is removed during Peak times. Figure 8a shows the results.

Figure 8 Cancelation rates for riders of different LGBT support and ethnicity

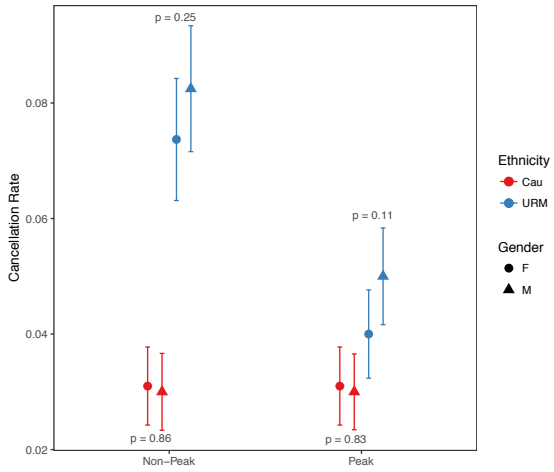


(a) Cancelation rates for riders of different LGBT support in different peak times.

(b) Driver cancelation rates for different LGBT support and ethnicities in different peak times.

Note. Cancelation rates are higher for URM riders and those that signal LGBT support during Non-Peak times. Bars show the 95% confidence interval.

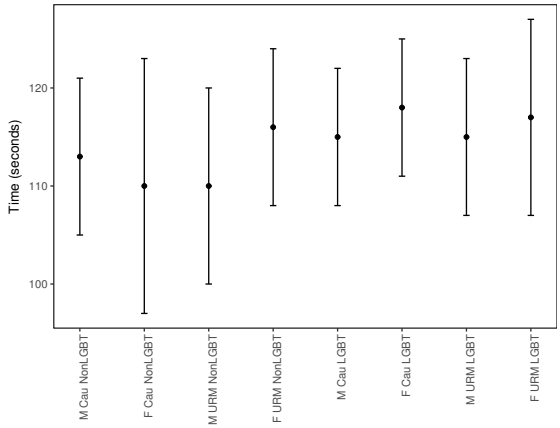
Figure 9 Driver cancelation rates for different rider ethnicities and genders



Note. Cancelation rates for riders of different ethnicities and genders at different peak times. Bars show the 95% confidence interval.

Figure 8b shows that a URM rider who also signals support for the LGBT community experiences higher cancelation rates than a URM rider that does not. Driver biases doubly impact these riders, resulting in cancelation rates that reach 8.7% during Non-Peak times.

Figure 10 Time between driver acceptance and driver cancelation, conditional on cancelation



Note. Drivers are canceling on riders in approximately the same amount of time. Bars are the length of one standard deviation due to being based on few observations.

Interestingly, we find no significant differences across Genders as shown in Figure 9. It is unclear why the biases previously found against women do not appear in our experiment. It could be that the platform change removed bias against women. Alternatively, the timing (during/post the #MeToo movement) may have played a role or the specific location of the experiment could be such that these biases did not previously exist.

One potential concern could be that the population of drivers is fundamentally different during Peak and Non-Peak times. Unfortunately, IRB restrictions prevent us from examining any driver characteristics. Nevertheless, we can observe the time between when a driver accepts a ride and the time at which they cancel a ride, conditional on canceling. Figure 10 shows that there are no differences across any treatments indicating that drivers are acting in a similar manner during Peak and Non-Peak times. We explore this more fully in the paper.

References

Bertrand, Marianne, Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *The American Economic Review* **94**(4) 991–1013.

Ge, Yanbo, Christopher R Knittel, Don MacKenzie, Stephen Zoepf. 2016. Racial and gender discrimination in transportation network companies.

Minear, Meredith, Denise C Park. 2004. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers* **36**(4) 630–633.