Zak Wegweiser
Oct 18 · 7 min read

AB TESTING AND EYE-TRACKING ANALYSIS

This project can be viewed at:
https://medium.com/@zak_wegweiser/introduction-de29cd533ca1

Introduction

The goal of this project was to conduct a series of tests to demonstrate the value of quantitative and qualitative user tests for evaluating different designs. As a group, Elvis Zhang, Rachel Wang, Zachary Espiritu, and I collected and analyzed user behavior through quantitative data in A/B Testing and qualitative data in Eye Tracking. We designed the following website for Memphis Taxis in two slightly different ways: (https://cs1300-ab-testing.herokuapp.com)

Memphis Taxis Reviews Services Contact

Where Travel Happens

If you're looking for a way to get away, a place where nothing can go wrong, the cheapest travel opportunities available, the most luxurious travel opportunities available, all of the travel opportunities available, or just looking for a taxi in Memphis, we've got you covered!

Find your ride

See what users are saying:

Steve Jobs

Professional Innovator

❝ Memphis Taxis is the most innovatative taxi service in the world.

★★★★★

Dean Baquet

Editor at *The New York Times*

❝ Wow, I wish we had something like this in New York.

★★★★½

Jeff Huang

Interface Extraordinare

❝ This taxi company has one of the best websites I've ever seen.

★★★★☆

The beginning of version A and B are the same

Ready for the ride of your life?

Yellow Cab	RideChange	Uber	Premier
Safer than most taxis	Cooler than most taxis	Cheaper than most taxis	Faster than most taxis
Rates start at $20	Rates start at $20	Rates start at $1	Rates start at $100
✔ Airport transport	✔ Cleaned hourly	✔ Flat rate	✔ Luxury travel
✔ Wheelchair-accessible	✔ Available 24/7	✔ Background checks	✔ Fast turnaround
✔ Metered fares	✔ Professional drivers	✔ Neighborhood drivers	✔ Personal chauffeur

Memphis Taxis

We get you to where you need to go.

Ready for the ride of your life?

Service	Starting Rates	Benefits			Link
Yellow Cab	$20	✔ Airport transport	✔ Wheelchair-accessible	✔ Metered fares	
Ride Change	$20	✔ Cleaned hourly	✔ Available 24/7	✔ Professional drivers	
Uber	$1	✔ Flat rate	✔ Background checks	✔ Neighborhood drivers	
Premier	$100	✔ Luxury travel	✔ Fast turnaround	✔ Personal chauffeur	

Memphis Taxis

We get you to where you need to go.

But the bottom of version A and B differ by display of content: vertical vs. horizontal

The users are introduced to the website with the same interface, but when they see the comparing price table the site changes. Version A consists of a vertical price table, while Version B consists of a horizontal one, as seen above.

Hypotheses

Null Hypotheses

The end goal was to determine if we could reject the same null hypothesis that **site A is no better than site B** based on each of the following metrics:

1. Click-through rate—the percentage of users that click on the website.
2. Time to click—the average time it takes a user to click.
3. Dwell time—the average time a user takes to return to the site after clicking.
4. Return rate—the percentage of users that return to the site after clicking off.

Alternative Hypotheses

For each null hypothesis, we created an alternative hypothesis, which is contrary to the null hypothesis. It is not possible to prove the alternative hypothesis, but we can try and make an assumption for why we would be able to reject our null hypothesis for each metric.

1. Click-through rate—A > B because most of the buttons are at the end of the page.
2. Time to click—B < A because the first call to action (CTA) is closer to the top of the page, so users may be inclined to click on each CTA as they scroll down the page.
3. Dwell time—B < A because the CTAs are in a vertical column, users will decide to review each website for themselves and then come back immediately to review the next website.

4. Return rate—B < A because of the same reason that the dwell time for B will be less than A. The layout of Version A will encourage users to review all of the options before settling on a single decision.

Eye Tracking Hypothesis

Version B will have a greater proportion of eye-gazes toward the left side of the screen than Version A because the important data for each of the comparisons are located on the left side of the screen.

. . .

A/B Testing

For each described metric, we individually computed the measurements and conducted the appropriate statistical test to determine whether or not we could reject our null hypotheses. My task was to choose between using a chi-squared test and a t-test. Both tests essentially compare the values from A and B to see how different they are from each other, but there are subtle differences.

T-tests are used when you are looking at the *means* of different data. This test indicates whether or not the mean in group A is significantly different to the to the mean in group B. Consequently, we use t-tests when analyzing the *average* time to click and *average* dwell time. This test will indicate whether or not those metrics will allow us to reject our null hypotheses.

Chi-squared tests are most commonly used when examining *categorical data*, such as the number of certain types of data, and they try to see whether the numbers are consistent with a null hypothesis. For this reason, we use chi-squared tests when analyzing click-through rate and return rate. This test will allow us to indicate whether or not we can reject the null hypotheses for those categories.

Computations: https://cs1300-stats-tests.herokuapp.com

I wrote up a PHP script to explain and compute the analyzations of each piece of data, which you can view here: https://cs1300-stats-tests.herokuapp.com. This site will allow you to click on which metric you want to analyze, and it will walk you through how it performs the correct statistical test.

The site thoroughly explains how these calculations work, but I will also give a very brief explanation here:

- Click-through and return rate—Compute the metric from the data logs. Then calculate the sum of all the $(O-E)^2/E$, where O is the observed value and E is the expected value, to get the chi-squared value. Compare this value with the probability value for 1 degree of freedom at 0.05 (which is 3.84). If the chi-squared value is greater than that, our data is statistically significant so we can reject the null hypothesis. Otherwise, we fail to reject it.
- Average dwell time and click time—Compute the average time from the data logs. Then use the sample size, sample mean, and standard deviation to compute a T value. Use a T-chart to find a critical T value in the 95% confidence interval (corresponding with our degrees of freedom). Compare our T value with the critical T value. If our T value is greater than the critical T value, our data is statistically significant, so we can reject the null hypothesis. Otherwise, we fail to reject it.

If all that math lingo was boring or confusing, don't worry. **The main result was we were unable to reject our null hypotheses for all our tests. This means our data was too similar to say one was better than the other.**

If you are interested in how this data was computed, you are welcome to view and manipulate the code here: https://github.com/zweg25/ab-testing.

. . .

Bayesian Probability

Another way to analyze these metrics is with Bayes' Beta Distribution Theorem. Essentially, it describes the certainty of our hypotheses

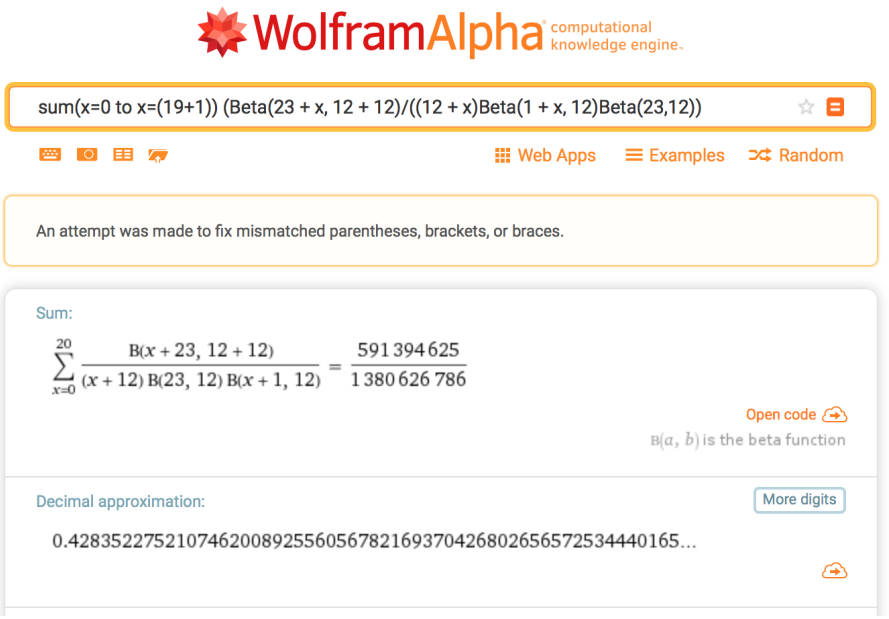
through probability. Below is an example of how to use a Bayesian A/B test for click-through rate, and the analysis of my data subsequently follows.

Bayesian A/B test for click-through

$$P(X > Y) = 1 - \sum_{j=0}^{c-1} \frac{B(a+j, b+d)}{(d+j)B(1+j, d)B(a, b)}$$

where
B(v, w) is a Beta Distribution (note that *B(1, 1)* is a uniform distribution)
a is the "number of clicks in X" + 1
b is the "number of non-clicks in X" + 1
c is the "number of clicks in Y" + 1
d is the "number of non-clicks in Y" + 1

How to compute Bayes' Probability for click-through rate



My Bayesian computation of click-through rate

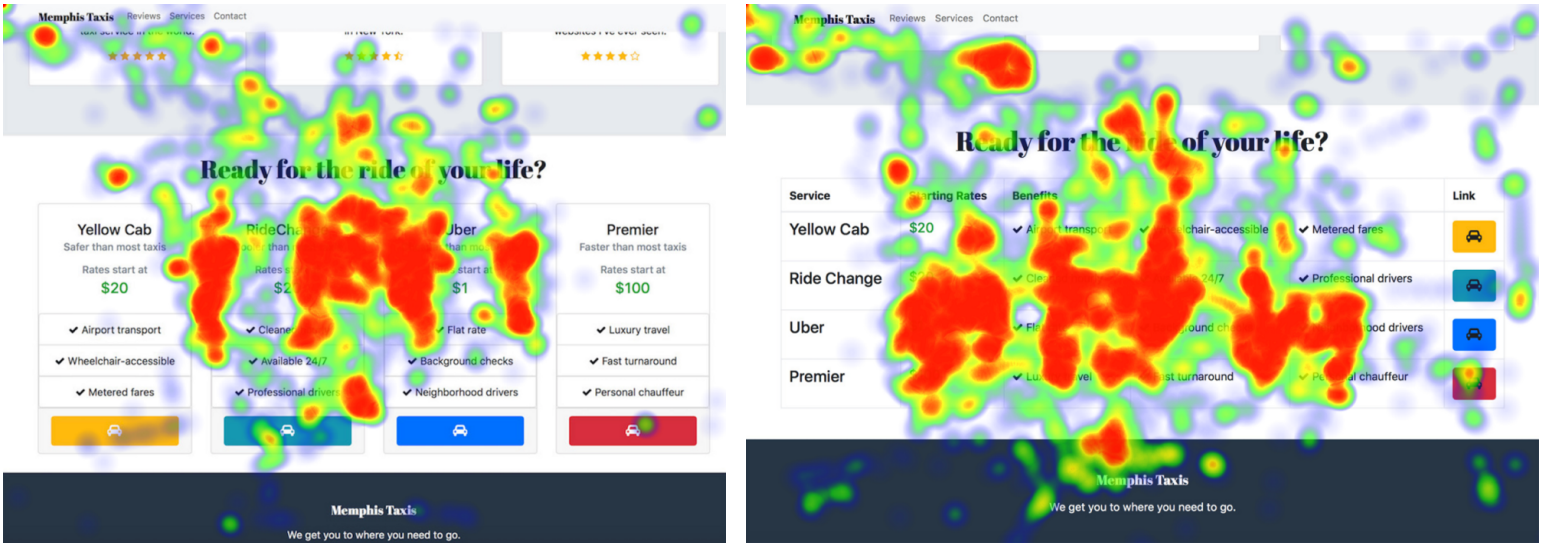
You can see this analyzation for yourself by inputting the following into WolframAlpha: *sum(x=0 to x=(19+1)) (Beta(23 + x, 12 + 12)/((12 + x)Beta(1 + x, 12)Beta(23,12))*

Since this project does not go very in-depth about Bayesian Probability Theorems, if you are interested, you can research more here: http://www.evanmiller.org/bayesian-ab-testing.html

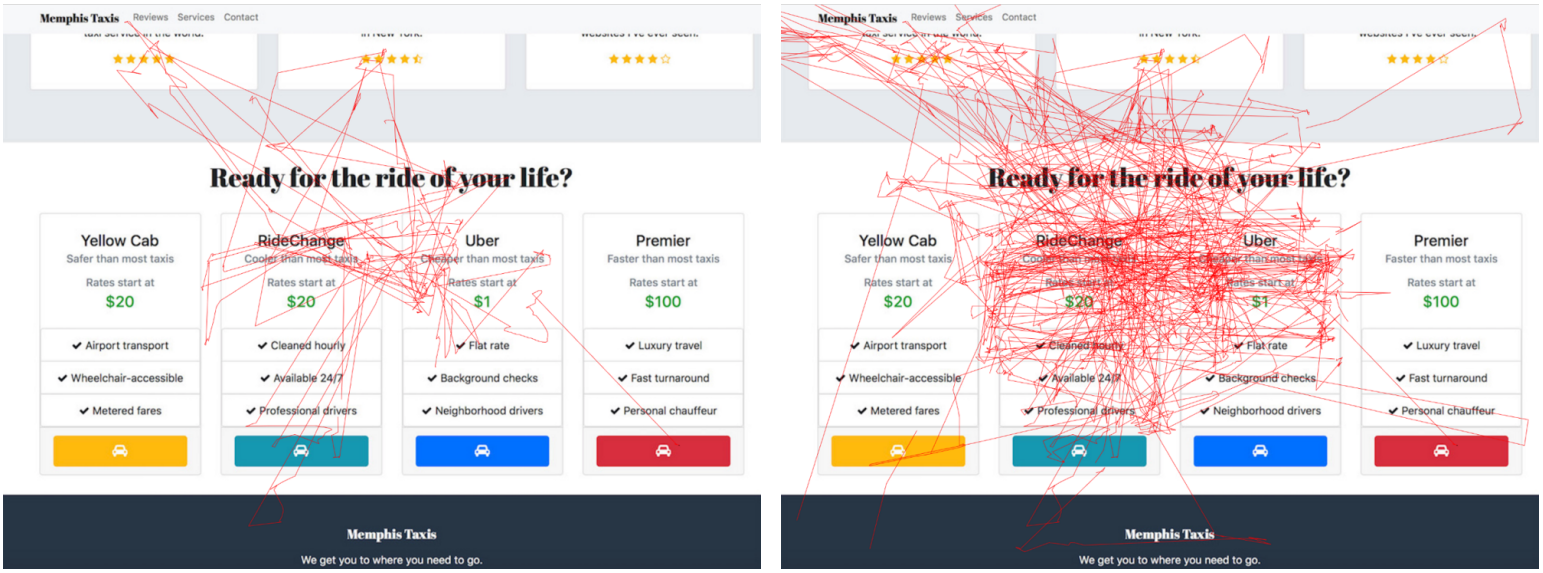
. . .

EYE-TRACKING

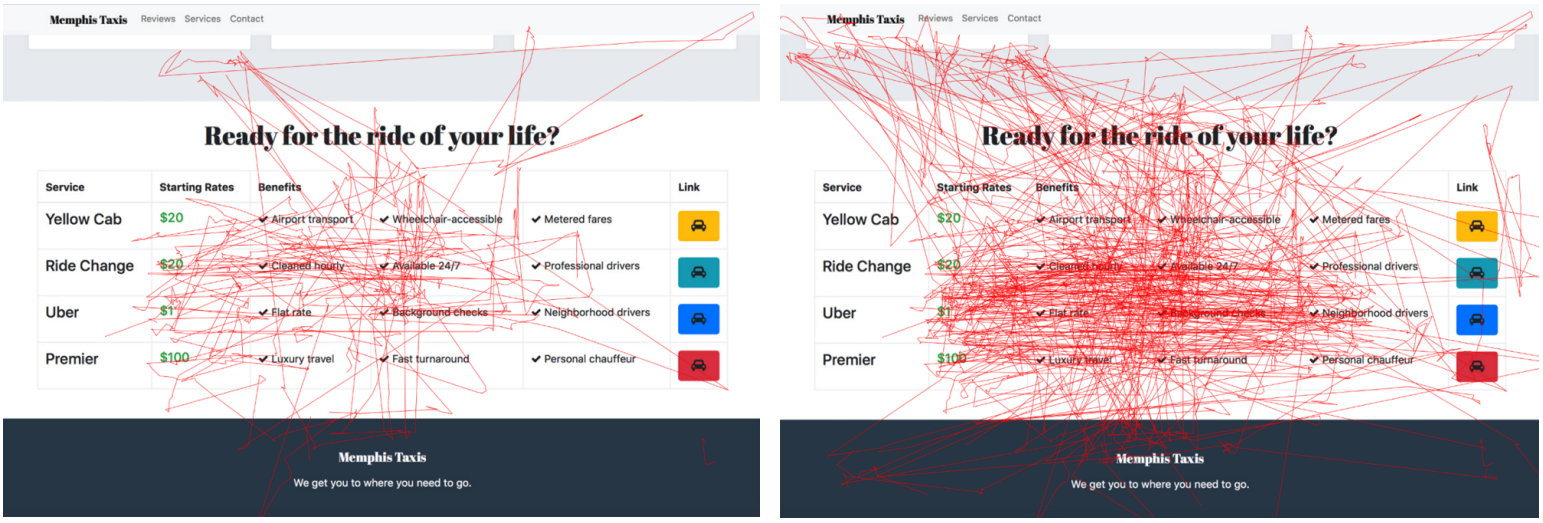
As mentioned, during this process we also used various eye-tracking equipment and software to watch two users' eyes as they explore our websites. After examining the eye-tracking logs, I created a script—viewable at the same website as above—to generate a heatmap and a replay of the users' eye-gazes:



Version A & Version B Eye-Tracking Heatmaps



Version A — Replay of User's Eye-Gazes



Version B— Replay of User's Eye-Gazes

Interpretation of Data

Looking back at our eye-tracking hypothesis, it seems that both users were actually more attracted to the middle of the screen. However, for Version A of our website this meant looking at just one or two car companies, but with Version B, the user actually read all the data displayed for every business. Although the data did not agree with our hypothesis, it generated valuable information about how users viewed our websites: they were attracted to the middle of the screen.

. . .

Conclusion

If we were conducting this experiment for a real taxi business in Memphis, we would not have a concrete answer for which version of the website to choose. Since we were unable to reject all of the null hypotheses throughout our A/B tests, I would consider leaning towards the eye-tracking test data to help make a decision. The Memphis Taxi company would have to decide if they want to direct users to one or two car companies like in Version A, or if they want users to read all the data before choosing like in Version B.

Reflecting on the tests as a whole, it is important to note that just because the A/B tests did not allow us to reject the null hypotheses does not make them useless. On the contrary, all the information gathered is useful to know before making a decision. However, a good learning experience is that A/B tests are better for comparing designs that have more differences, rather than subtle differences. In those cases, A/B tests would most likely provide more significant data than the eye-tracking experiments because the result might directly indicate which version of the website is better and how. But, in this case, the eye-tracking data had more of an impact on our results because it was better at comparing the subtle change we implemented.

As a whole, both tests are useful for comparing two versions of a website, but eye-tracking is better at testing subtler modifications, while A/B testing is more useful when comparing more significant changes.