

Project Report

A/B Testing and User Testing: Memphis Taxis



Part 1: A/B Testing

“A picture says a thousand words”, but is that too many? The following A/B tests examine the possibility that a bad picture is worse than no picture at all. We’ve created two versions of the same website. Version A contains the exact same content as Version B, except without the vague, generic taxi pictures. Fundamentally, the testing asks a question about user priorities: do people care more about the visual presentation of a page or the quality of the content?

Consider our null hypothesis, that there is no conceivable difference in any of the following four parameters given between Version A and Version B of the websites. If our test hypothesis is correct, we would expect the parameters to change as described in the table below (Fig. 1).

	Null Hypothesis	Alternative Hypothesis
Click Rate	There is no statistically significant difference in click rate between A and B	There is a statistically significant difference between the click rates of A and B.
Time to Click	There is no statistically significant difference in time to click between A and B	The time to click of A will be higher than B. (People will spend more time looking at the content instead of glancing at pictures.)
Dwell Time	There is no statistically significant difference in dwell time between A and B	The dwell time of A will be higher than B. (People will have picked the right link on the first go.)
Return Rate	There is no statistically significant difference in return rate between A and B	There is a statistically significant difference between the return rate of A and B.

Fig. 1 Table displays null and alternative hypothesis for each of the four parameters.

Heroku Link: <https://safe-shelf-75338.herokuapp.com/>

Version A:

Memphis Taxis: Version A

This page contains information about taxi and cab companies in Memphis, Tennessee.
Not endorsed by any of the companies listed, or the city of Memphis.

Reserve with YellowCab Taxis

Safe, Reliable Taxi Services from Yellow Cab and Checker Cab are available in the Memphis Metro area 24 hours a day.

Reserve with RideCharge taxi

Our drivers are the most professional drivers in the industry. Our drivers are licensed and required to successfully complete a formal training.

Reserve with Memphis Uber

We are so much cheaper than taxis..!


Reserve with Premier taxi

Our goal is to efficiently transport our passengers in a safe, polite and timely manner at a fair price.

Version B:


Memphis Taxis Version B

This page contains information about taxi and cab companies in Memphis, Tennessee.
Not endorsed by any of the companies listed, or the city of Memphis.




Learn More

Safe, Reliable Taxi Services from Yellow Cab and Checker Cab are available in the Memphis Metro area 24 hours a day.




Learn More

Our drivers are the most professional drivers in the industry. Our drivers are licensed and required to successfully complete a formal training.



Learn More

We are so much cheaper than taxis..!



Learn More

Our goal is to efficiently transport our passengers in a safe, polite and timely manner at a fair price.

The metrics are calculated in the table below:

	Definition	Version A	Version B
Click Rate	Proportion of sessions with at least one click	42.4%	61.9%
Avg. Time to Click	Time from page load to that first click	15.51 secs	32.02 secs
Avg. Dwell Time	Time spent on external page before return	34.62 secs	6.36 secs
Return Rate	Proportion of sessions that returned	33.33%	19.04%

Fig 2: Metrics for Version A and Version B. See raw values in Appendix A.

Note on calculations: When users reloaded the page upon returning, the page sometimes reloaded as the other version, since the page version is determined randomly. Thus, every reload to a different version of the page was counted as a new user. Summary statistics without this approximation are attached in Appendix B.

Formulas for Calculations: The following formulas were used to compute the above metrics.

$$\text{Click Rate} = \frac{\text{number of first clicks}}{\text{number of unique users}}$$

$$Avg\ Time\ to\ Click = \frac{\sum_{\forall\ unique\ users} click\ time - load\ time}{number\ of\ first\ clicks}$$

$$Avg\ Dwell\ Time = \frac{\sum_{\forall\ unique\ users} last\ click\ time - load\ time}{number\ of\ returning\ users}$$

$$Return\ Rate = \frac{number\ of\ returning\ users}{number\ of\ unique\ users}$$

Statistical Tests:

I chose to use a Chi-Squared Test to analyze the click rate and return rate I am interested in the “success rate” (that is, the rate of clicking and returning) for each of the statistic. I used an one-tailed independent samples T-test for the average time to click and dwell time since I am interested in the difference in means between the two versions for those two specific statistics.

Click Rate:

$$X^2 = \sum_{all\ cases} \frac{(observed - expected)^2}{expected}$$

Observed	Click	No Click	Total
Version A	14	19	33
Version B	13	8	21
Total	27	27	54

Fig. 3 Observed values for the click rate

Expected	Click	No Click	Total
Version A	16.5	16.5	33
Version B	10.5	10.5	21
Total	27	27	54

Fig. 4 Expected values for the click rate

$$X^2 = \frac{(14 - 16.5)^2}{16.5} + \frac{(19 - 16.5)^2}{16.5} + \frac{(13 - 10.5)^2}{10.5} + \frac{(8 - 10.5)^2}{10.5}$$

$$= 0.38 + 0.38 + 0.6 + 0.6 = 1.9481$$

There are 2 degrees of freedom, and the p-value is 0.1628. Thus, the result is not significant at $p < 0.05$.

Average Time to Click:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}}$$

$$\bar{X}_1 = 15510.286 \text{ ms}$$

$$\bar{X}_2 = 32017.385 \text{ ms}$$

$$N_1 = 14, N_2 = 13$$

$$s_1 = 20812.711, s_2 = 48003.824$$

$$\begin{aligned} t &= \frac{15510.286 \text{ ms} - 32017.385 \text{ ms}}{\sqrt{\frac{(14 - 1)20812.71^2 + (13 - 1)48003.82^2}{14 + 13 - 2} \left(\frac{1}{14} + \frac{1}{13} \right)}} \\ &= \frac{-16507.0989}{\sqrt{\frac{(13)20812.71^2 + (12)48003.82^2}{25} (.148)}} = -1.17 \end{aligned}$$

There are 25 degrees of freedom, and the p-value is 0.1256. The confidence interval is calculated as follows:

$$se = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \left(\frac{1}{N_1} + \frac{1}{N_2} \right)} = 14429.125$$

$$\begin{aligned} \text{Confidence interval} &= -16507.0989 \pm 0.1256 * 14429.125 \\ &= [-14694.8008, -18319.397] \end{aligned}$$

Thus, the result is not significant at $p < 0.05$.

Average Dwell Time:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}}$$

$$\bar{X}_1 = 32618 \text{ ms}$$

$$\bar{X}_2 = 6356.25 \text{ ms}$$

$$N_1 = 11, N_2 = 4$$

$$s_1 = 63050.547, s_2 = 1888.289$$

$$t = \frac{32618 \text{ ms} - 6356.25 \text{ ms}}{\sqrt{\frac{(11-1)63050.547^2 + (4-1)1888.282^2}{11+4-2} \left(\frac{1}{11} + \frac{1}{4}\right)}} = \frac{28261.75 \text{ ms}}{\sqrt{\frac{(10)63050.547^2 + (3)1888.282^2}{13} \left(\frac{1}{11} + \frac{1}{4}\right)}} = 0.58803$$

There are 11 degrees of freedom, and the p-value is 0.83. The confidence interval is calculated as follows:

$$se = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = 28303.625$$

$$\text{Confidence interval} = 28261.75 \text{ ms} \pm 0.83 * 28303.625$$

$$= [4769.74, 51753.76]$$

Thus, the result is not significant at $p < 0.05$.

Return Rate:

$$\chi^2 = \sum_{\text{all cases}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Observed	Return	No Return	Total
Version A	11	22	33
Version B	4	17	21
Total	15	39	54

Fig. 5 Observed values for the return rate

Expected	Return	No Return	Total
Version A	9.17	23.83	33
Version B	5.83	15.17	21
Total	15	39	54

Fig. 6 Expected values for the return rate

$$\chi^2 = \frac{(11 - 9.17)^2}{9.17} + \frac{(22 - 23.83)^2}{23.83} + \frac{(4 - 5.83)^2}{5.83} + \frac{(17 - 15.17)^2}{15.17}$$

$$= 0.37 + 0.14 + 0.58 + 0.22 = 1.3055$$

There are 2 degrees of freedom, and the p-value is 0.2532; thus, the result is not significant at $p < 0.05$.

Results:

The conclusions from A/B Testing are summarized below.

	Test Type	p-Value	Verdict
Click Rate	Chi-Squared Test	0.16	Failed to reject
Avg. Time to Click	Independent Samples T-Test	0.13	Failed to reject
Avg. Dwell Time	Independent Samples T-Test	0.83	Failed to reject
Return Rate	Chi-Squared Test	0.25	Failed to reject

Fig 6: We failed to find a statistical difference between the two versions of the website for any of the metrics.

Part 2: User Testing

In Part 1, I attempted to find the optimal way to communicate information about taxi companies to a user. In this part, I wanted to focus on what specific information people cared the most about when booking a taxi.

Hypothesis:

Users' are primarily concerned about price when booking a cab; all other factors that influence the choice of a cab are secondary to that primary concern.

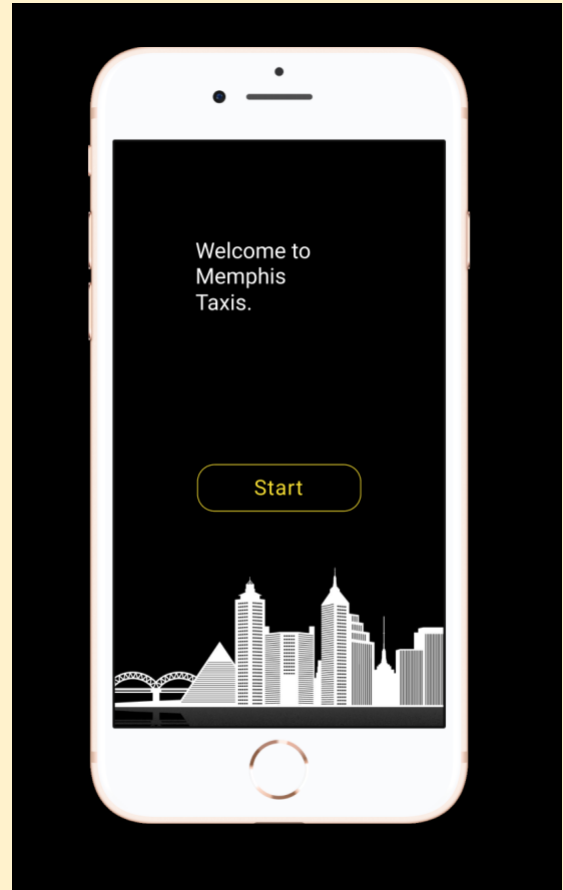
Prototype

User Testing Instructions:

Before interacting with the app, users were asked to complete two tasks:

1. Describe what they thought they could accomplish on the app.
2. Verbalize their primary considerations when booking a taxi.

Users were then instructed to attempt to book a taxi given their verbalized considerations.

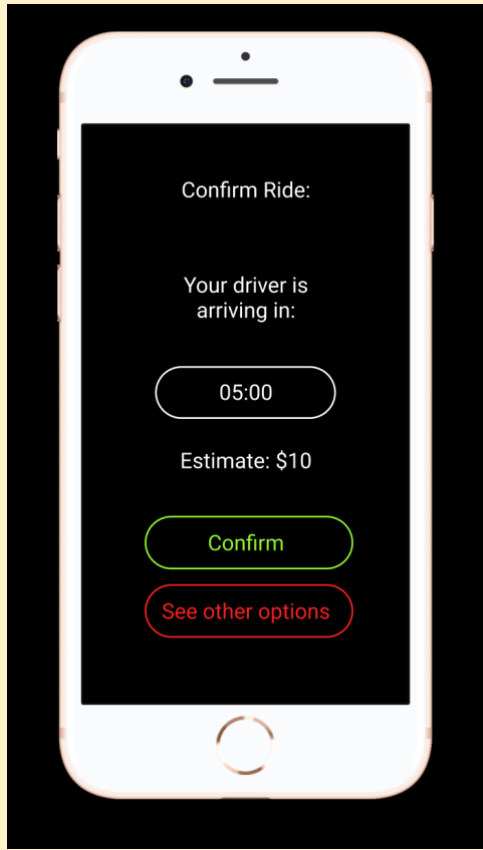


Results:

Our user testing seemed to confirm our hypothesis. All three of the users verbalized cost as one of their primary concerns. The third user specifically wrote that cost was her primary concern and that she appreciated that cost was the first filter.

	Average Time	Completion Rate	Error Rate
App Functionality	45 secs	100%	0%
Considerations	100.3 secs	100%	0%

In general, it seemed that users who had a clearer idea of their priorities completed both tasks in the same amount time. Since there were no errors and all users successfully completed the task, it is apparent that the navigation and usability of the app seemed reasonable.



Most users wished that they had more information about their ride before they booked it. In addition to more driver information and ride information, users also wanted to be able to estimate the cost of their ride and have ways to confirm what kind of vehicle their driver was in for safety reasons. Additionally, users wished that there was more confirmation before the ride was booked, citing the fact that they felt like the transition to the “booked” page was rather sudden.

Aesthetically, some users felt intimidated by the dark scheme of the app and wished the design was friendlier and had more color.

Although User Testing was much less statistically rigorous and quantitatively satisfying, the open ended-ness of the testing allowed users to present considerations that I had not previously thought necessary or even considered. In the future, I would change the questions to be more specifically related more specific sub-tasks. Additionally, I would do more rounds of user testing in order to make changes to the design of the app and re-evaluate.

Part 3: Conclusions

Recommendations to Memphis Taxis:

Although the results from A/B Testing were fairly inconclusive, User Testing produced some actionable results. In general, all users wanted more information about their ride than was provided. Memphis Taxis should consider adding the following features:

1. Ability to view driver information (make and model of car, picture of driver, license plate)
2. More ride information (number of seats available, current driver location, time to destination, time until driver arrival, estimated cost)
3. Give users the ability to cancel a ride

Furthermore, some users felt negatively about the dark aesthetics of the app and wished that there was more color in the app in order to make it more approachable.

Testing Methods:

Although both User Testing and AB Testing require designers to evaluate a hypothesis of some kind. User Testing allows for much more open-ended responses, which may reveal considerations that the designer hadn't considered before. On the other hand, A/B Testing provides a much more rigorous standard of proof that can distinguish effectively the subconscious effects of two relatively similar designs.

In future designs, it may be helpful to employ User Testing in order to do more exploratory hypothesis testing and use A/B Testing to gain design specificity after the general direction of the design has been determined by User Testing.

Appendix A: Python Code (Version A)

```

#####
@title - Filter AB Testing results
@author - Karen T
@date - Nov 1, 2018
#####

import sys

#read file
fileName = sys.argv[1]
textFile = open(fileName, "r")
#pattern = textFile.readlines()[0]

def process(file):
    sessions = {}
    # (unique users, click rate, time to click, dwell time, return rate)
    summaryStats = {'A': [0, 0, 0, 0, 0], 'B': [0, 0, 0, 0, 0]}
    for line in textFile:
        wordList = line.split(" ")
        timestamp = wordList[0]
        version = wordList[3]
        loadTime = wordList[4]
        clickTime = wordList[5]
        clickId = wordList[6]
        userId = wordList[7]

        if userId in sessions: #consequent action
            if sessions[userId][0] == 0: #first click
                if int(clickTime) != 0 and int(loadTime) != 0:
                    summaryStats[version][1] += 1
                    summaryStats[version][2] += (int(clickTime) - int(loadTime))
                sessions[userId][0] = 1
                sessions[userId][1] = int(clickTime)
            elif sessions[userId][0] == 1: #return
                if int(sessions[userId][1]) != 0:
                    summaryStats[version][3] += (int(loadTime) - int(sessions[userId][1]))
                    summaryStats[version][4] += 1
                del sessions[userId] # reload is considered a new user

        else: #First page load
            if version == "A":
                summaryStats['A'][0] += 1
            else:
                summaryStats['B'][0] += 1
            sessions[userId] = [0, 0] #on site
    print len(sessions)
    print "sessions"
    print sessions
    print "summary stats"
    print summaryStats

process(textFile)

```

Results:

[version]: [unique users], [click rate], [total time to click], [total dwell time], [return rate]
'A': [33, 14, [41369, 8954, 1473, 27119, 4590, 631, 76155, 12568, 3866, 9455, 17300, 1783, 8547, 3334], [5442, -8954, 4588, 4630, 315909, 7151, 5710, 2615, 4351, 5024, 34332], 11]
'B': [21, 13, [761, 35265, 148619, 4742, 7656, 2167, 3387, 1190, 2144, 20253, 76690, 7544, 105808], [7784, 6020, 4951, 6670], 4]

Appendix B: Python Code (Version B)

```

#####
@title - Filter AB Testing results
@author - Karen T
@date - Nov 1, 2018
#####

import sys

#read file
fileName = sys.argv[1]
textFile = open(fileName, "r")
#pattern = textFile.readlines()[0]

def process(file):
    sessions = {}
    # (unique users, click rate, time to click, dwell time, return rate)
    summaryStats = {'A': [0, 0, 0, 0, 0], 'B': [0, 0, 0, 0, 0]}
    for line in textFile:
        wordList = line.split(" ")
        timestamp = wordList[0]
        version = wordList[3]
        loadTime = wordList[4]
        clickTime = wordList[5]
        clickId = wordList[6]
        userId = wordList[7]

        if userId in sessions: #consequent action
            if sessions[userId][0] == 0: #first click
                if int(clickTime) != 0 and int(loadTime) != 0:
                    summaryStats[version][1] += 1
                    summaryStats[version][2] += (int(clickTime) - int(loadTime))
                    sessions[userId][0] = 1
                    sessions[userId][1] = int(clickTime)
            elif sessions[userId][0] == 1: #return
                if int(sessions[userId][1]) != 0:
                    summaryStats[version][3] += (int(loadTime) - int(sessions[userId][1]))
                    summaryStats[version][4] += 1
                #del sessions[userId] # reload is considered a new user

        else: #First page load
            if version == "A":
                summaryStats['A'][0] += 1
            else:
                summaryStats['B'][0] += 1
            sessions[userId] = [0, 0] #on site
    print len(sessions)
    print "sessions"
    print sessions
    print "summary stats"
    print summaryStats

process(textFile)

```

Results:




[version]: [unique users], [click rate], [total time to click], [total dwell time], [return rate]

'A': [18, 12, 213179, 346466, 10]

'B': [12, 10, 307513, 10971, 2]

Appendix C: UserTesting.Com

<input checked="" type="checkbox"/>	☆	UserTesting Support	COURSES-B/CS1300	Your UserTesting video (3/3) is complete: Taxi App Test - Use...	Oct 30
<input checked="" type="checkbox"/>	☆	UserTesting Support	COURSES-B/CS1300	Your UserTesting video (2/3) is complete: Taxi App Test - Use...	Oct 30
<input checked="" type="checkbox"/>	☆	UserTesting Support	COURSES-B/CS1300	Your UserTesting video (1/3) is complete: Taxi App Test - Use...	Oct 30
<input checked="" type="checkbox"/>	☆	UserTesting Support	COURSES-B/CS1300	Your Order with UserTesting - Usertesting logo" width="188" h...	Oct 30

Participants using computers					
<input type="checkbox"/>		kierant92 ★★★★★ 25 - Male - \$100,000 - \$124,999 - Australia 1 Note 0 Clips See Answers	<div>Summarize this session</div>		
<input type="checkbox"/>		Tanko1314 ★★★★★ 29 - Male - \$80,000 - \$99,999 - Australia 8 Notes 0 Clips See Answers	<div>Summarize this session</div>		
<input type="checkbox"/>		elliebelly0307 ★★★★★ 21 - Female - \$40,000 - \$59,999 - Australia 8 Notes 0 Clips See Answers	<div>Summarize this session</div>		