Testing Web App Design with A/B Testing

Liam Fruzyna
liam.fruzyna@marquette.edu
Marquette University
Milwaukee, WI, USA

Reid Holben
reid.holben@marquette.edu
Marquette University
Milwaukee, WI, USA

Vidit Kalani
vidit.kalani@marquette.edu
Marquette University
Milwaukee, WI, USA

ABSTRACT

We sought to test how users of different types interacted with a website displaying public transit stations. Subjects were asked to identify stations in different economic areas, while we also tracked their interactions based on their handedness, age, and more general demographic information. The results were largely inconclusive, showing that different demographics behave largely the same.

KEYWORDS

datasets, public transit, income, handedness, demographics

ACM Reference Format:
Liam Fruzyna, Reid Holben, and Vidit Kalani. 2019. Testing Web App Design with A/B Testing. In ,. ACM, New York, NY, USA, 7 pages.

INTRODUCTION

This paper aimed to examine how different demographic groups interact with web features. Public transit was chosen because of people's relatively familiarity with it as well as the ease of displaying the data.[14] The project utilized Amazon Mechanical Turk to collect user input. Respondents followed a series of instructions for interacting with the web app. After using the web app respondents completed a survey to inform us about their interactions and themselves.

Background on Data

The public transit data was found in two different data sets from the Oak Ridge National Laboratory. *Public Transit Stations*[9] provided information on fixed rail transit stations in the United States. *Public Transit Routes*[10] provided information on fixed rail transit routes in the United States, most importantly which stations belong to which routes. Both were found on the website of the Homeland Infrastructure Foundation. We also used the 2016 IRS *SOI Tax Stats*[13] data set to determine income levels of certain areas. These data were merged together into one uniform data set for the ease of our use.

METHODS

Data Preparation

Our data was acquired from three separate data sets, two from the HIFLD and one from the IRS SOI Tax Stats. These data sets needed to be combined into one cohesive data set, where each row represented a single station. We utilized R with, as well as, the dplyr library to clean the data. Since the final data set was to be based on individual stations we started with the *Public Transit Stations* data set and removed any unnecessary columns and correctly formatted zip codes as strings. For the *SOI Tax Stats* data, sums were totalled up for each zip code which were then used to compute the average reported income of each zip code. Each zip code was then given a bracket it would fit in based on its average income. This bracket and average income were appended to the data for each station. Finally, the *Routes and Stations* data was also used to append which routes each station belonged to.

Data Description

Our final data set consists of 4505 public transit stations, each with the following features:

- STATION: The name of the station.
- STA_ID: A unique identifier for each station.
- SYSTEM: The name of the transit system the station belongs to.
- LATITUDE and LONGITUDE
- STATUS: Regarding if the station is open.

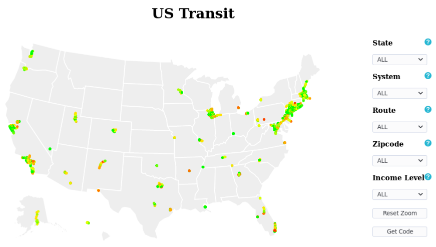

Figure 1: The web app after page load.

- CITY, STATE, and ZIP
- STATIONSYS: A combination of the station and system names.
- income: Average reported income of the zip code.
- bracket: Income level bracket (1-10, >10).
- RTE_NAME: A double-bar ("||") separated list of routes the station belongs to.

Web App

Our web app was developed using the D3 JavaScript library. It consists of two main components, a map of the United States and a column of options for filtering the set of displayed stations. The map displays colored dots representing public transit stations, with the color of the dot representing the income bracket the station falls in. The dots can be hovered over to display the station name and average reported income. These dots can be filtered using a variety of options displayed along the side of the map. Selecting a state, system, route, or zip code will zoom the map in on that selected group of points. States on the map can also be clicked on to zoom in and the zoom can be reset by pressing the "Reset Zoom" button on the sidebar. Finally at the bottom of the sidebar, a "Get Code" button is provided for MTurk respondents to get a metadata code about their usage, which is collected in our survey.

A/B Testing Methodology

We had three null hypothesises that we wanted to be answered from our A/B testing. First, age has no effect on how many times states on the map are clicked, versus how many times the states are selected from the drop down. The number of clicks on both states on the map, as well as, the state drop down menu were counted while the user was using the web app. To compare the results to the age of the participants, either value was used in a one-way ANOVA with the age range of the participant on the other side.

$$\text{Number of Clicks} = \beta_0 + \beta_1(\text{Age Category of Participant})$$

We predicted that the null hypothesis would be rejected and that despite clicking less overall, older users would click on the map at a lower proportion than younger users. It has been seen that older people have slower response times[3] and computer skills became more and more difficult to build[6] which may cause them to spend less time exploring the web app as they are spending more time completing the required tasks.

Second, handedness has no effect on a user's preference for the side a sidebar is displayed on. In order to test a preference for the sidebar, the side that the sidebar appears on was randomized and recorded. To quantify the user's preference the user was walked through a series of steps, clicking on a point on the map, then immediately clicking on a drop down menu on the sidebar. The time to

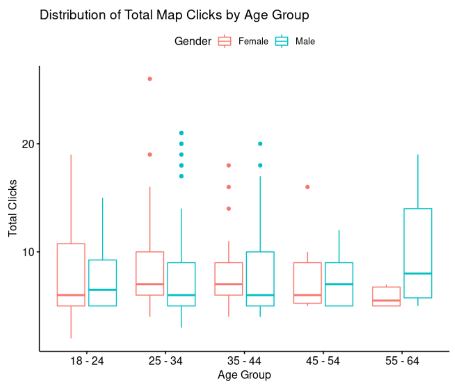

Figure 2: Box plots of total number of clicks across the map and drop down menu for various ages and genders.

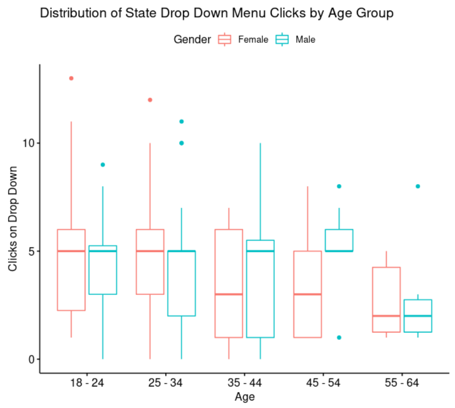

Figure 3: Box plots of the number of clicks on the drop down menu for various ages and genders.

complete this task was measured. A two-way ANOVA was used to compare the results to both the user's handedness and the side the sidebar appeared on.

$$\text{Time to Complete} = \beta_0 + \beta_1(\text{Handedness of Participant}) + \beta_2(\text{Sidebar Side})$$

We predicted that the null hypothesis would be confirmed, that the user's handedness does not affect their time to reach the sidebar on either given side. However, we do expect to see a slight preference to the left because in most of the world people read left to right. The opposite situation was analyzed by Damien M. Berahzer who looked at vertical scroll relocation. He found the scroll bar being on the left also made interactions take longer, even with individuals rating both systems with the same level of difficulty[1]. People expect the scroll bar to be on the right due to conventions so we expect a preference to the left for the sidebar due to conventions. We can assume neither left or right handed people will be overall different because Goodin found that response time does not vary between left and right handed people[4]. Research has also found that in other tasks like a sport such as handball, sex also has some relation to performance which shows this is another factor we should consider[5].

Third, education has no effect on how a user interprets the relative poorness of an area. To determine the user's interpretation, the user was guided to a specific route on the map and asked how many stations they believed were in poor areas. This number was recorded in the survey and compared to the user's education level in a one-way ANOVA.

$$\text{Number of Poor Stations} = \beta_0 + \beta_1(\text{Education Level of Participant})$$

We predicted that the null hypothesis would be rejected and users of a higher education level would determine that fewer stations were in poorer areas. Public transit accessibility is a key issue as cities urbanize, and identifying stops that are in poorer areas often finds public transit stops that are underfunded[8]. Different visualization designs for map data have previously changed people's interpretation of the data [12]. If people of higher income who are examining this data have difficulty seeing which areas are poorer, a gap in providing help to the areas may be created.

Survey

The survey was designed in three parts to test each of our hypotheses. In general, the participants would interact with our web app then report their usage to the survey. Afterward using the web app, we would collect general demographic data on the users. The first part was interested in the user's intuition for interacting with the map. The participants were instructed to select three different states without instructing them how to. The number of clicks on states and the state drop down menu were individually recorded while they did this. The second part was interested in the user's preference for a sidebar on the left or right of the map. The participants were given three tasks again, this time to select a point on the map, which started a timer, then choose a new option from the sidebar, which

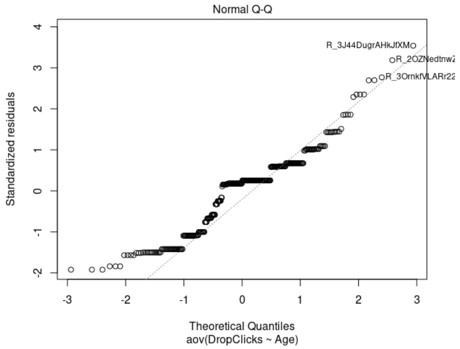

Figure 4: QQ Plot for the model of number of drop down clicks to age.

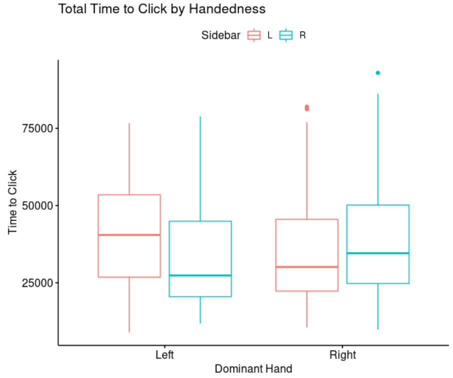

Figure 5: Box plots of total time to click for left and right handed people for both sidebar sides.

ended the timer. These three times would be used to determine if the user has a preference. Finally, in our third part, we guided the participants to a specific route on the map and asked them how many of the stations on the route they believed were in poor areas. This would be used to determine how the users viewed the data presented to them.

RESULTS

We first compared the total number of clicks on the map and state drop down menu. From our initial box plot this was not promising (Figure 2). There did not appear to be any connection between the clicks, age, or gender.

Next, we compared the number of clicks on the state drop down menu to the age of the participants. An initial box plot showed potential (Figure 3). The one-way ANOVA resulted in a p-value of 0.0115. This appeared significant so we moved onto Tukey's HSD test. Unfortunately none of the category pairs proved individually significant enough. This was further confirmed with a very poor QQ Plot (Figure 4). We failed to reject our null hypothesis. There did not show to be a correlation between age and number of clicks or how they clicked.

Our second comparison was between handedness and time to reach the random sidebar. Due to large times we saw while our data was being collected we made an adjustment to our web app to tell the user the next step when they clicked on one of the stations that starts a timer. We compared our results from before and after this change and found in general times improved. On average times improved by 6.5 seconds for the three tests, from 18.6 seconds to 12.1 seconds, a 35% improvement. Due to this improvement, only submissions after this change were analyzed. When comparing the sidebar side and handedness with the box plot produced appeared to show a sign of a correlation (Figure 5). However, when the ANOVA was computed none of the predictors were significant. This meant we failed to reject our null hypothesis, however, that is what we expected. We did not see an overall preference to the left side though.

We did find a correlation between the third time test and the side of the sidebar. It tested significant with a p-value of 0.0493 and was confirmed with a Tukey HSD test returning a p-value of 0.0493 as well. But when the QQ Plot was made it was very skewed (Figure 6).

Our final comparison was between the number of reported poor stations and the education level of the participant (Figure 7). This one-way ANOVA was very insignificant resulting in a p-value of 0.688. Again, we failed to reject the null hypothesis. No connection was found between education and how our web app was interpreted.

CONCLUSION

As seen in the results, our data didn't provide very useful insights. This was primarily due to the inconsistent results of our survey. There were, however, some lessons learned about A/B testing in

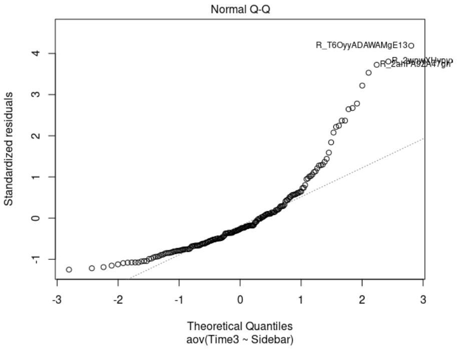

Figure 6: QQ Plot of the model of the third time test and the sidebar side.

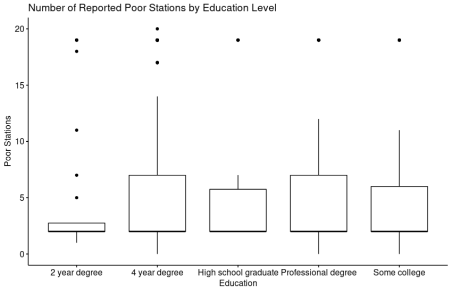

Figure 7: Box plots of number of reported poor stations for various education levels.

general. First with distribution to Mechanical Turk. We discovered early in our testing that increasing the reward increased the quality and quantity of submissions.[2] Due to the length of our survey we would have likely been better off further increasing the reward of our survey if nothing else was changed. This would likely result in less overall data, but the data would be of an improved quality.

We were also able to increase the quality of our results by improving our instructions. It has been found that "Mechanical Turk is best suited for tasks in which there is a bona fide answer"[7] so respondents don't have to guess and check. We found biggest improvement was from adding instructions into our web app so respondents did not have to leave the page to find the next instruction. As mentioned previously this resulted in a 1/3 improvement in times recorded on the web app. We do recognize that our instructions needed further improvement. At the end of our survey there was only about a 50% completion rate. We believe that making our list of tasks shorter and easier would improve the quality of our results. This would include removing time-based tasks which we did not find to be successful and requiring less precise clicks on the web app which slowed down respondents.

The responses we received from the survey questions about the web app didn't seem to follow any kind of pattern. Many results appeared to be from bots or lazy workers, which is a common problem on Mechanical Turk[11]. Adding an attention check to our survey questions may help determine which answers are meaningful and which are spam. Being able to remove poor submissions would improve the quality of data we could use in our analysis.

APPENDICES

- Our git repository can be found on GitHub at: https://github.com/mail929/COSC4500-Project
- Our web app can be found at: https://mail929.github.io/COSC4500-Project/webpage/

REFERENCES

[1] Damien M Berahzer. 2005. Scroll Placement and Handedness. (Apr 2005). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.138.3505&rep=rep1&type=pdf

[2] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5. https://doi.org/10.1177/1745691610393980 arXiv:https://doi.org/10.1177/1745691610393980 PMID: 26162106.

[3] Neil Charness, Patricia Holley, Jeffrey Feddon, and Tiffany Jastrzembski. 2004. Light Pen Use and Practice Minimize Age and Hand Performance Differences in Pointing Tasks. *Human Factors* 46, 3 (2004), 373–384. https://doi.org/10.1518/hfes.46.3.373.50396 arXiv:https://doi.org/10.1518/hfes.46.3.373.50396 PMID: 15573539.

[4] T. A. Ortiz R. S. Chequer D. S. Goodin, M. J. Aminoff. 1996. Response times and handedness in simple reaction-time tasks. (1996).

[5] SENOL DANE and ALI ERZURUMLUOGLU. 2003. SEX AND HANDEDNESS DIFFERENCES IN EYE-HAND VISUAL REACTION TIMES IN HANDBALL PLAYERS. *International Journal of Neuroscience* 113, 7 (2003), 923–929. https://doi.org/10.1080/00207450390220367 arXiv:https://doi.org/10.1080/00207450390220367 PMID: 12881185.

[6] Katharina V. Echt, Roger W. Morrell, and Denise C. Park. 1998. EFFECTS OF AGE AND TRAINING FORMATS ON BASIC COMPUTER SKILL ACQUISITION IN OLDER ADULTS. *Educational Gerontology* 24, 1 (1998), 3–25. https://doi.org/10.1080/0360127980240101 arXiv:https://doi.org/10.1080/0360127980240101

[7] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[8] Megan Campbell Cynthia L. Bennett Vicki Le Sean Pannella Robert Moore Kelly Minckler Rochelle H. Ng Jon E. Kotaro Hara, Shiri Azenkot. 2015. Improving Public Transit Accessibility for Blind Riders by Crowdsourcing Bus Stop Landmark Locations with Google Street View: An Extended Analysis. (2015). https://dl.acm.org/citation.cfm?id=2717513

[9] Oak Ridge National Laboratory. 2017. Public Transit Stations. data retrieved from World Development Indicators, https://hifld-geoplatform.opendata.arcgis.com/datasets/public-transit-stations.

[10] Oak Ridge National Laboratory. 2017. Public Transit Stations. data retrieved from World Development Indicators, https://hifld-geoplatform.opendata.arcgis.com/datasets/public-transit-routes.

[11] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (01 Mar 2012), 1–23. https://doi.org/10.3758/s13428-011-0124-6

[12] Cobourov Saket, Scheidegger and Borner. [n.d.]. Map-based Visualizations Increase Recall Accuracy of Data. ([n. d.]). https://doi.org/10.1111/cgf.12656

[13] Internal Revenue Service. 2016. SOI Tax Stats - Individual Income Tax Statistics - 2016 ZIP Code Data (SOI). data retrieved from World Development Indicators, https://hifld-geoplatform.opendata.arcgis.com/datasets/public-transit-routes.

[14] Vanessa Smith. 2019. Safer Stops: Increasing Public Perceptions of Safety through Bus Stop Design in the City of Greater Sudbury. (Apr 2019). https://qspace.library.queensu.ca/handle/1974/26138