Emily Grace Lobel and Sophia Grace Trump
Professor Dougherty
cs215
18 May 2018

SCIENTIFIC JOURNAL
B L U B U S: A New Application

LINK TO GOOGLE DOCUMENT:
https://docs.google.com/document/d/1SadZzNm96EieLFa3Avle6Jv0b_JJcOav5fQbb2yWtxo/edit?usp=sharing

Intro

For our application, we wanted to focus on usability for people in the Bi-College Community. We felt that this was wise as we could relate to this community and feel what needed solving within it. In deciding what would actually be useful for people within the Bi-Co, we thought of the idea to create a new application to keep track of the Blue Bus schedule.

The current Blue Bus website, while functional, is not very accessible. The URL is unmemorable and does not contain the word “blue bus,” so unless one bookmarks it the website is hard to get to in the first place. The text is small and the page allows no manipulation of the buses; just a list of every bus that goes to and from each college. We wanted our application to quickly provide the same information, but with more usability. We wanted to allow for a simple interface that quickly allows the user to find the most relevant information, as well as set alarms and keep busses/alarms that are important readily accessible within the application. Once we had sketched out our interfaces, we set upon this task.

Workflow (experimental design)

Our experimental design was inspired by A/B testing (DTUI 156). The question we wished to examine within our study was: Which interface is more usable (i.e., provides the best user experience), the Tabs interface or the Dropdowns interface?

We define a usable interface as one that offers high speed of performance, low error rate, high success rate, and high subjective satisfaction (DTUI 14, 1). Usability of interfaces was tested using a combination of qualitative and quantitative measures. The qualitative measures aimed to measure subjective satisfaction. The quantitative measures aimed to measure speed, error rate, success rate, and subjective satisfaction.

Based on previous discussions in the course about the value of simplistic design, we hypothesized that the Tabs interface would be more usable than the Dropdowns interface. As such, our null hypothesis is that Tabs is no more usable than Dropdowns.

In order to control for the confounding variable of interface presentation order, we randomly assigned participants to two groups (DTUI 155-156, 2). Group A was presented first with Tabs followed by Dropdowns. Group B was presented first with Dropdowns followed by Tabs and had a sample size of one.

Our overall test structure was as follows:

1. Participants were presented with a pre-survey.
2. Participants completed a series of timed tests we deemed relevant to the interfaces through a combination of requirements and task analysis (DTUI 106-108, 3).
3. Participants were presented with a post-survey, in which they wrote down difficulties had with each interface (and suggestions for improvement), inspired by the think aloud approach (DTUI 152.154).

Within the pre-survey, we asked questions to gain demographic data from participants. We also asked questions to gauge prior experience with the BiCo Blue Bus.

The timed tasks were recorded using a combination of a screen recorder and a stopwatch.

The timed tasks requested of participants follows:

1. Go to the official blue bus page.
2. Find when the next bus will be leaving from Bryn Mawr (tell this to the experimenter).
3. Find when the next bus will be leaving from Haverford (tell this to the experimenter).
4. Adjust the preset alarm settings to set alarms 30 minutes before bus departure.
5. Set an alarm for the bus leaving from Bryn Mawr on Tuesday, 8:50 AM.
6. Find the alarm you just input.
7. Set an alarm for the bus leaving from Haverford on Friday, 11:10 PM.
8. Find the alarm you just input.
9. Set a custom alarm from Bryn Mawr on Monday at 4:17 PM.
10. Set a custom alarm from Haverford on Wednesday at 12:33 AM.
11. Find the alarms you just input.

The post-survey was inspired by the System Usability Scale, and thus sought to gauge participant impressions of the interfaces post-task analysis (4). Participants were also asked questions specific to Tabs and Dropdowns, such as which interface the participant preferred, and if participants would use either interface again. Participants were also asked for suggestions for improvement for both interfaces, which were used in the construction of our third interface.

Results / Analysis

Taken from our experimental design, our data can be partitioned into two parts: pre-survey data, which provides qualitative insight to the background of our test population, and task and post-survey data, which together provide qualitative and quantitative data about UX during the experiments. ***[View the data for each trial here. All data visualizations can be found in the "Data Visualizations" section at the end of the report.]***

From the pre-survey data, 100% of participants were Bryn Mawr students. 55.6% were sophomores, 33.3% were juniors, and 11.1% were seniors ***[See Chart A]***. 100% of participants had taken the BiCo Blue Bus before. Frequency of Blue Bus usage ranged from 1 time per week to once every two months, with the majority of the population reporting Blue Bus usage more than 3 times per week and 2 times per week ***[See Chart B]***. 44.4% of participants reported that they did not like the official Blue Bus website, with 33.3% reporting they had no feelings about the official website ***[See Chart C]***. 44.4% of participants reported they miss the Blue Bus monthly. 100% of participants reported using the official Blue Bus website to stay on schedule with the Blue Bus ***[See Chart D]***.

From task and post-survey data, we define **speed** as the amount of time it takes to complete a task, measured in seconds. The speed of each task is the arithmetic mean of speed across all participants. We treat each task individually as they are not necessarily performed in sequence (7). We use relative score of average speed per task to compare speed between interfaces (7). *[View the speed data per task, including relative scores, here.]*

The Tabs interface had an average speed of 19.2 seconds per trial (SD 17.2). The Dropdowns interface had an average speed of 29.2 seconds per trial (SD 20.8). Using the geometric mean across trials for the relative speed scores of the Tabs and Dropdowns interfaces, Tabs had a combined relative increase in task speed of 69.9% and Dropdowns had a combined relative decrease in task speed of 40.5%. A two tailed t test using the speed across tasks yielded a p value of 0.0913, which is not quite statistically significant. This suggests that we cannot reject the null hypothesis, and that there is no significant difference between the interfaces ***[See Chart E]***.

We define **number of errors** for each task as the number of mouse clicks that do not lead to task completion. Each task's error rate is the arithmetic mean of errors across all participants. We use relative scores to compare number of errors between interfaces. *[View the error data for each task, including relative scores, here.]*

The Tabs interface had an average of 1.6 errors per task (SD 1.9). Dropdowns had an average of 1.9 errors per task (SD 1.9). Taking the geometric mean across trials for the relative error scores of the Tabs and Dropdowns interfaces, Tabs had a combined relative decrease in errors of 38.4% and Dropdowns had a combined relative increase in errors of 27.6%. A two tailed t test using the number of errors across tasks yielded a p value of 0.6056, which is not statistically significant. This suggests that we cannot reject the null hypothesis, and that there is no significant difference between the interfaces ***[See Chart F]***.

We define **success rate** as a variation on the time-based efficiency metric (5), by taking into account error rate in task completion across all tasks (as inspired by the success rate metric

here: 12). We take error into account in our efficiency metric by reducing the task completion variable, which ranges from 1 (successfully completed) to 0 (unsuccessfully completed), by 0.1 per error (i.e, if ten errors are made, the task is counted in the metric as unsuccessfully completed). *[View the performance data for each task here.]*

The Tabs interface had an average success rate of 0.2 goals/second (SD 0.1). Dropdowns had an average success rate of 0.1 goals/second (SD 0.1). Taking the geometric mean across trials for the relative error scores of the Tabs and Dropdowns interfaces, Tabs had a combined relative increase in success rate of 31.8% and Dropdowns had a combined relative decrease in success rate of 46.6%. A two tailed t test using the success rate across tasks yielded a p value of 0.0116, which is statistically significant. This suggests that the null hypothesis can be rejected, and that the Tabs interface had a higher success rate than Dropdowns ***[See Chart G]***.

We define **subjective satisfaction** as the mean of SUS scores, based on post-survey responses. [*View the SUS scores per participant here.]*

Tabs had an average SUS score of 27.7 (SD 2.6). Dropdowns had an average SUS score of 27.7 (SD 2.4). Taking the geometric mean across trials for the relative error scores of the Tabs and Dropdowns interfaces, Tabs had a combined decrease of 0.5% in SUS score. Dropdowns had a combined increase of 0.5% in SUS score. A two tailed t test using the SUS scores across tasks yielded a p value of 0.9159, which is not statistically significant. This suggests that we cannot reject the null hypothesis, that there is no significant difference between the interfaces, suggesting that Tabs and Dropdowns have no statistically significant difference in terms of subjective satisfaction ***[See Chart H]***.

However, the results of the post-survey should be noted, as they provide a more qualitative perspective. 55.6% of participants reported that they would not use the Dropdowns interface again ***[See Chart I]***, but 77.8% of participants reported that they would use the Tabs interface again ***[See Chart J]***. 77.8% of participants reported that they preferred the Tabs interface over the Dropdowns interface ***[See Chart K]***.

We define **usability** as the geometric mean of the relative scores of speed, number of errors, success rate, and subjective satisfaction (1).

According to our metric, the Tabs interface is 32.5% more usable than the Dropdowns interface. The Dropdowns interface is 30.7% less usable than the Tabs interface. However, as 3 out of the 4 multiplicands in our metric were deemed statistically insignificant, we cannot reject the null hypothesis in terms of usability. This suggests our data shows no significant quantitative difference between the usability of Tabs and Dropdowns ***[See Chart L]***.

However, as noted in the subjective satisfaction section of our report, the majority of participants indicated a preference for the Tabs interface over the Dropdowns interface, and that they would use the Tabs interface again (but not Dropdowns). Qualitatively, therefore, our findings support the findings of the usability metric, suggesting that Tabs is in fact more usable than Dropdowns, which confirms our initial hypothesis.

Mistakes

We made many mistakes over the course of this project, one of them arguably being underestimating the scope of what we were taking on. Neither of us had much experience with Javascript, HTML, or CSS, and assumed that because we were only working with two real functions (setting a pre-set alarm and setting a custom alarm), the project would be fairly light. However, upon actually going about coding these functions, we discovered just how complicated this would turn out to be. Furthermore, we wanted our website to provide a usable but familiar interface, so we looked at the actual code of the official Blue Bus website, which was convoluted and overcomplicated. Parsing through it was a formidable challenge, let alone implementing it properly into our website.

As time went on, we also became used to the way our website looked and worked. Once we introduced other people to it, we realized that we had gotten so caught up in our terminologies and vocabulary (literal and technical) that it had become too complicated and was lost on others. Since both of our interfaces were so complicated, this resulted in a rather extensive task analysis required, which then led to a more strenuous testing experience, which is reflected in our data. The complexity of the interfaces made it difficult for participants to learn and adjust to them during experimentation. We attempted, in coding our third interface, to bring humanity and usability back into our website while keeping everything we had worked on.

What we learned

We learned a lot during this process, from coding the basics of HTML to understanding how other people interact with a website. We firstly learned not to underestimate the power and the complexity of Javascript. When coding our two interfaces, we realized that some Javascript and HTML had become so complex in both their individuality and their linked together-ness that changing them, as we had initially planned to do, would set us back another week. In order to achieve everything we had to set out to do, we left some plans in the dust, prioritizing making a working website.

During testing, we learned even more. What had become natural and comprehensible to us quickly became too complex for others to understand within a short amount of time (as a good website should be able to be understood within a short amount of time), and we quickly had to make adjustments to fix this. We learned, firstly, that testing is hard. It is difficult not only to sit and watch someone struggle with your product, but also to analyze the data that comes from it. We also had some errors with our interfaces, including instances where the interface would stop working if we didn't refresh it, so that was frustrating. However, we got through it and we were able to learn a lot from the post-survey results; we did our best to take peoples' suggestions. By watching others struggle with our designs, improving the interfaces became not only a requirement for the project, but a desire to make the experience better for those who use them.

As mentioned above, the "Tabs" interface was the most popular interface -- the one people felt most confident using. In the post-survey, most people advocated for more color and more "graphical components." We wanted to keep our website within the theme of blue/white/black as the blue buses themselves look, but for a splash of color we added yellow and red, Bryn Mawr and Haverford's respective colors (along with black). We kept in mind low

vision and color blind users by making sure to add black outlines to colored components. We added more flair to the already existing components of the website. We made the graphic elements easier to see and appreciate by making them bigger and adding more whitespace. Since people had appreciated the multiple functionalities of the website despite its complexity, we aimed to make our website simpler. We added more feedback so that the user can know that functions are working as they intended; for example, the custom alarms button now gives a confirmation that the alarm was set. We added a “help” button as well as more text explaining functions. We are aware that this does not assist or speak to the intuitiveness of the website, but we wanted to keep all of the functions we had worked so hard on, and have settled for this alternative for now. We look forward to further improving the website in the future. We also changed much of the already existing text on the website to hopefully clarify exactly what each page and function does. This speaks to the fact that we had become too self-involved and, in the future, need to remember how much humanity and usability counts, even in the development stages of the website.

The main thing we hope to improve upon in the future is to integrate more flexibility in our website. For example, by adding a “color free” option, or a way to adjust font sizes. Though we feel we made our interface more intuitive post testing, we hope to make the interface even more intuitive, though we admit that we would need to run another round of testing to determine what steps would need to be taken to achieve this. Perhaps including more than one preset alarm setting, or having the the alarms appear in a more visually pleasing way under the “My Alarms” section would aid in achieving this. Furthermore, we know that at times buttons stop working for nebulous reasons, so we would like to improve that. Overall, we are proud of what we have accomplished and are excited to continue improving our interface in the future. We feel very good about the finished product and are very excited to have gone from complete novices to somewhat experienced web developers.

Data Visualizations

Chart A

Chart B

Chart C

Chart D

Chart E

Chart F

Chart G

Chart H

Chart I

Chart J

Chart K

Chart L

Articles read (sources)

https://www.nngroup.com/articles/usability-metrics/
https://isps.yale.edu/node/16697
https://www.usability.gov/how-to-and-tools/methods/task-analysis.html
https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html
https://www.w3schools.com/
https://vwo.com/ab-testing/
http://www.brynmawr.edu/transportation/bico.shtml