Design and Prototyping Project

TOOBA SHEIKH, 101028915
GARRISON SU, 101232418
MERRAJ MASSTAN, 101186611
ARDA DOLANAY, 101113511

Authors' addresses: Tooba Sheikh, 101028915, toobasheikh@cmail.carleton.ca; Garrison Su, 101232418, GarrisonSu@cmail.carleton.ca; Merraj Masstan, 101186611, MerrajMasstan@cmail.carleton.ca; Arda Dolanay, 101113511, ArdaDolanay@cmail.carleton.ca.

Contents

1 USER STUDY DESIGN

1.1 Study Design

In the previous assignment, we created a product called QuizCreate, which is a quiz-making website mainly for educational use targeted for students and teachers. Now that we are imagining that QuizCreate has been released to the public, we want to find out if our website's features accommodate the target users' needs for them to not only be able to improve their studying routine, but to also check if the website's features provide support and are effective.

For this user study, we have two research questions: (1) Is there a difference in the effectiveness of the website between users of different age demographics? (Specifically looking at users in high school, undergraduate, graduate, and teachers/professors' categories.) (2) How satisfied are the users with the accessibility of the interface and the features it offers compared to a known study/quiz website such as Quizlet? These will help us receive valuable feedback from users who all have various levels of experience and education, while also discovering faults in our product as well as missing important features that we haven't considered.

For this study, we have two independent variables: the division of age demographics, which is slightly related to the users' current status (high-school, undergrad, and alumni/postgraduate students, as well as teachers), and the application that participants will be using; our website, QuizCreate, and Quizlet, a similar web application that primarily uses flashcards as a study technique. The advantage is that we would obtain proper feedback relative to not only the participants' respective levels of education, but their computer skills as well. For the latter, according to Nielsen Norman Group, only 5% of the population between the ages 16 to 65 across 33 countries has high computer-related skills [1]. Moreover, between the ages 25 and 60, people's ability to use websites declines by 0.8% per year [2], proving that there is a significant variation of skills between age groups. For users under the age of 16, they study in different settings where external resources are not provided, meaning that they are not encouraged (yet not prohibited) to use applications like Quizlet. However, the user study will not be able to measure our dependent variables with specific areas of education, such as mathematics and literature, or even with different university programs, such as law, computer science, and more. Our dependent variables are the efficiency of use (which includes time), user satisfaction, and perceived ease of use. With these variables, we would get data that can represent the effectiveness of our product's functionalities and measure the users' experience. Yet, the downside of this selection of variables is that we cannot assume that all users will have the same exact experience level for a pure comparison between each of them. In other words, there is a greater possibility of outliers when we obtain the data.

Furthermore, our user study will be conducted remotely through the web. This is because we want users to have the freedom to do their given tasks in the most natural setting depending on their personal preferences; either they prefer studying at home, at a library, etc. Also, it will allow some flexibility to perform their tasks at any time during the day within the timeline of the user study. Sadly, since this is being conducted through the web, we are sacrificing the ability to directly observe the users, and there is a risk of technical difficulties that may rise, such as the possibility of a power outage or the potential unavailability of internet connection, and we must prepare in advance to face and fix. In addition, the user study only has one control condition, which is the use of an already available quiz-making website, specifically Quizlet. However, there is no control condition for the age demographics independent variable because there is no default age that we can test.

Finally, since we have two independent variables, the design of our user study will be both between-subjects and within-subjects depending on the variable. On one hand, for the various age demographics, we will be using a between-subjects design; each participant belongs to a certain specific age group. On the other hand, for the applications that

participants will use, namely Quizlet and QuizCreate, we will use a within-subjects design; each group defined by age will perform the same tasks in both applications.

1.2 Data Collection

One of the most important features in QuizCreate is the ability to create a quiz, but to do so in an efficient way as well as providing a feeling of support for the user. By capturing the amount of time users take to create a quiz, the results will tell us if the website allows the user to create quizzes very quickly or if it is a lengthier process. To obtain results for this data, we will be tracing timestamps from the time they first enter the quiz creation page up until they have finished creating the quiz at the very end. This functionality will be temporarily set up in the back end of the website only for the purpose of this step in the user study. The dependent variable that the results will help us measure is the application's efficiency of use. This data is **objective-quantitative** since these are numerical measurements.

Next, we have data that is similar to the previous one, which is the time searching for a specific quiz. The data will help us to find out if our product's search feature is helpful and responsive enough for our users. Once again, we will be capturing the timestamps as well to obtain these results, which will help us measure the user efficiency of use. Since the data captures time with no regard to any opinion or preference, the data is **objective-quantitative**.

The third quantitative data that we will be obtaining is how the user felt supported throughout their experience learning the features of the product. This question will ask the participants of our user study to rate their experience based on this criterion on a scale from 1 to 10, therefore this data is **subjective-quantitative**. This is a question that will be asked in a questionnaire that we plan on providing to participants after they have performed their tasks on our website throughout the study. The dependent variable relating to this data is the user's satisfaction.

Furthermore, since our product is a website, then it is naturally available using various devices. In our questionnaire, we will be asking what type of device and browser they have used; whether the participant used a computer, a mobile phone, a tablet, etc., and if they were on Google Chrome, Safari, Firefox, etc. This is related to a few other questions to be asked in the questionnaire, which is the user's opinion for each of the website's features, namely creating a quiz, taking a quiz, and searching for a quiz. With both of these data, we will be able to correlate them and find out on which device the product needs to improve, as well as the specific features that are affected. The dependent variable that these two results will help measure is the perceived ease of use. The data for the type of device and browser is **objective-qualitative** since it is categorical and is not based on opinion. This will be captured through screen recordings alongside comments that support what they have experienced while using the website. The data for the participant's opinion on the main features is, of course, **subjective-qualitative**. It will be acquired during the questionnaire.

Finally, the user's overall opinion of the website's visual design will be picked up through the questionnaire. With this information written in text form, we can get input on some visual features on various aspects, such as the color scheme, the menu bar (including its position on the page and the links attached to it), the icons (both on the menu bar and on the search/home page), and more. This data will help us measure two dependent variables; user satisfaction and the application's efficiency of use. Of course, it is also **subjective-qualitative** since these non-numerical results are based on many participants' perspectives and ideas.

1.3 Analysis Plan

As mentioned in the section before, we will be collecting six types of data from the users as described in the previous section. We will be analyzing each data type separately to get a better understanding of the results individually. This

will be followed by comparing the results from the analysis, to see if the data provides any evidence of the following research questions:

- Is there a difference in the effectiveness of the website between users of different age demographics? Specifically looking at users in high school, undergraduate, graduate, and teachers/professors' categories
- How satisfied are the users with the accessibility of the interface and the features it offers compared to a known study/quiz website such as Quizlet?

For context purposes, the following table show the participants will be in the following groups in the user study:

Independent Variables	Highschool Students	Undergraduate Students	Graduate Students	Professors/Teachers /Lecturers
QuizCreate	Group 1-1 Week 1	Group 2-1 Week 2	Group 3-1 Week 3	Group 3-1 Week 4
Quizlet	Group 1-2 Week 5	Group 2-2 Week 6	Group 3-2 Week 7	Group 3-2 Week 8

Fig. 1. Table of Participation Groups based on Independent Variable

This table will be used in the analysis plan to define the groups are being referred to.

Quantitative Data:

- **Time spent on creating a quiz:**
  The hypothesis or the question trying to be answered here is that: The longer it takes the user to create the quiz, the worse the usability. This is relevant to both research questions as this can show discrepancies between age groups as well as usability between the two software applications. The participants will be given a set of questions that they will use it to create the quiz. The data will be analysed as follows:
  - **Descriptive Statistics:**
    Mean, mode and median of the time taken for each test group. This will provide a general idea of the central tendency of our data distribution.
  - **Graph:**
    - A boxplot will also be developed between each group because this will really help compare the results for the different datasets. This will provide meaningful information regarding the differences in the mean and the lower quartiles for each group.
    - A histogram will also be developed for the data between QuizCreate and Quizlet for each age group. Example: Group 1-1 will be compared to Group 1-2. This will show if the data is in the normal distribution for each website and will highlight if the which website took longer to use. And it will help ensure that the age differences (high school vs undergraduate) do not affect the results
  - **Inferential Statistics:**
    This data will be the ratio data as it measures time. If the data distribution results in being a normal

distribution, then parametric tests will be used. Otherwise, non-parametric test will be used. As there are more than two groups for age vs age and this is between subjects:

* If normal: ANOVA test
* If not normal: Kruskai-Wallis test

As there are only two groups for Quizlet vs QuizCreate and this is within subjects:

* If normal: Paired T-test
* If not normal: Wilcoxon test

The hope is that the results will show that there is consistency in time usage between each age group and that it takes less time to create a quiz in QuizCreate than it does in Quizlet. This will answer the hypothesis: "The longer it takes the user to create the quiz, the worse the usability" in terms of both of the research question.

- **Time spent on searching for a quiz**:
  The hypothesis or the question trying to be answered here is that: The longer it takes the user to search for a quiz, the worse the quality of the QuizCreate search and matching algorithm that displays the results. This is relevant to only one research question regarding the Quizlet vs QuizCreate. This does not need to test the difference between age groups, as searching the system for a specific quiz should be consistent between participants. This test should also be very minimally impacted by the visual display of the quizzes because the quizzes are displayed as simple list after the search. The only environmental variable that can impact the search time is the type of device and browser used as a slower system would result in a slower result. The main data, time, is measured from the search till the user clicks on the quiz they were looking for. The participants will be also asked about the device and browser they used in the study. The data will be analysed as follows:
  - **Descriptive Statistics:**
    Mean, mode and median of the time taken for Quizlet vs QuizCreate group. The standard deviation will be focused on as well, to view any range of data. The age division will be eliminated, so all the data sets for age will be combined. This will again provide a general idea of the central tendency of our data distribution.
  - **Graph:**
    * A histogram will also be developed for the time data between QuizCreate and Quizlet. This will show if the data is in the normal distribution for each website and will highlight if the which website took longer to provide results.
    * A clustered bar graph data of each device group and browser group will be created to show if the type of device/browser impacted the result in any meaningful way by visually comparing them. So, the X-axis would be the different age groups and for each age group, multiple bars will represent the different devices and browsers
    * The type of device/browser impact will be discussed in the qualitative data section. It was mentioned here as it might have a major impact which must be considered while analysing this particular data set.
  - **Inferential Statistics:**
    The main data, which is the time to finish creating a quiz will be the ratio data as it is measured in time. If the data distribution results in being a normal distribution, then parametric tests will be used. Otherwise,

non-parametric test will be used. As there only two groups for Quizlet vs QuizCreate and the study is within subjects:

* If normal: Paired T-test
* If not normal: Wilcoxon test

The results will ideally show that our software QuizCreate is either faster or equal to Quizlet. Any discrepancies in the search results could ideally be linked to the device/browser.

- **How the user felt supported throughout their experience learning the features of the product**
  The hypothesis or the question trying to be answered here is that: The higher the satisfaction rating, the better the quality of support the product provides. This is relevant to both research questions as this can show any issues occurring between age groups as well as usability between the two software applications. The data from the questionnaires will be analysed as follows:
  - **Descriptive Statistics:**
    Mean, mode and median of the time taken for each test group. This will again provide a general idea of the central tendency of our data distribution. The standard deviation will be focused on for each group as there might be a higher deviation the younger or older groups.
  - **Graph:** This is pretty similar to the first quantitative data analysis as this is the best way to analyze the data between these data sets.
    * A boxplot will also be developed between each group because this will really help compare the results for the different datasets. This will provide meaningful information regarding the differences in the mean and the lower quartiles for each group
    * A histogram will also be developed for the data between QuizCreate and Quizlet for each age group. Example: Group 1-1 will be compared to Group 1-2. This will show if the data is in the normal distribution for each website and will highlight which website took longer to use. And it will help ensure that the age differences (high school vs graduate) do not affect the results.
  - **Inferential Statistics:**
    This data will be ordinal data as it is measures 1 - 10. Therefore, non-parametric tests will be used. As there are more than two groups for age vs age and this is between subjects:
    * Kruskai-Wallis test

    As there only two groups for Quizlet vs QuizCreate and this is within subjects:
    * Wilcoxon test

    The results will hopefully show a higher rating in support for QuizCreate vs Quizlet. The difference in age groups will hopefully show no discrepancy in feeling supported between older and younger groups.
    .

Qualitative Data:

- **Type of device, and browser**:
  This data will be collected using questionnaires. This is an important dataset to analyze, due to the vast variety of devices and browser used. This could potentially point out what compatibility issues might need to be considered while continuing development. The analysis techniques being used would be Grounded Theory. This was chosen as the data could result in a number of theories about the devices/browser used. As mentioned

earlier, a bar graph to compare the affect on the time taken would be the most important data to be considered. This would be done on the date sets for the age demographic variable, as analyzing the quizlet website for device efficiency is not needed. The question to be answered here would be:

- Does the type of device affect the efficiency of use as well as their satisfaction?

The steps would be as follows:

- Open coding to find meaningful code list for time taken on each feature, browser, and device.
- Axial coding to categorize these codes.
- Selective coding to find any relationships.
- This will be followed by trying to build a theory based on the relationship found.

While there are no themes that are already created, the hope is that the themes found would point the issue towards devices and not the website itself, so that compatibility can be worked as an outside requirement rather than needing to change the code for each browser/device.

- **User opinion for the website's main features: creating a quiz, taking a quiz, searching a quiz - Screen-Recording**

  This data will be collected screen-recording. The analysis techniques being used would be Inductive Thematic Analysis. Inductive analysis was used because we don't want to impose themes on the set of data, and want themes emerge regarding how the users feel about the features. The question being answered would be:

  - Does the user feel like they have enough tools and that they have enough control to do what they want through the use of the given features?

  The steps would be as follows:

  - Re-watch the screen-recording to familiarize with the data.
  - Assign preliminary codes to main actions and words found in the screen recording.
  - Find patterns or themes in the codes
  - Review, define and name themes.

  Looking at this data would hopefully result in positive themes about the satisfaction and ease of use of the website.

- **User's satisfaction on the overall design of the website:**

  This data will be collected using a questionnaire. The analysis techniques being used would be Deductive Thematic Analysis. Deductive analysis was used because the software was designed to have a positive impact on the user. The question being answered would be:

  - Is the product's aesthetics related to the user's sense of satisfaction? (do they want to use the product more?)

  The steps would be as follows:

  - Get accustomed to the data by creating creating graphs.
  - Assign preliminary codes to connections found in the data.
  - Find patterns or themes in the codes
  - Review, define and name themes.

  - Looking at this data would hopefully result in positive themes about the user satisfaction.

1.4 Participants

The participant demographic for the study will encompass a wide range of individuals linked within the educational institutions, including both students and teachers. This eligibility criteria will be carefully structured to target specific age groups and educational levels, ensuring that the study's findings are relevant and applicable to its intended audience.

The study will include participants ranging from high school students aged 14 to 17, who are navigating the critical years of teenage turmoil, all through the 9th grade to 12th grade, to mature students aged 18 to 29, who might still be suffering from the very prevalent protagonist syndrome seen in today's times, in higher education or pursuing advanced studies. Additionally, educators aged 30 and above are included. We shall acknowledge their pivotal role in the educational process and their unique perspective on the effectiveness of educational tools.

To ensure the study's focus remains on active members of the educational community, anyone not currently enrolled in or employed by an educational institution is excluded. This exclusion criterion is designed to avoid the participation of casual users, thereby maintaining the study's relevance to its target demographic.

The recruitment strategy will be twofold, encompassing both online and on-site methods. Online recruitment leverages the power of social media platforms, educational forums, and email newsletters to reach a broad audience of potential participants. By partnering with educational institutions, the study leverages existing networks to circulate information, thereby enhancing its reach and credibility.

On-site recruitment involves direct engagement with potential participants at schools, colleges, and educational fairs. This approach will benefit from face-to-face interactions, providing an opportunity to convey detailed information about the study's objectives and address any questions or concerns in real-time.

To encourage participation, the study will offer incentives such as free access to premium features of the website, additional educational resources, or certificates of participation. These incentives are thoughtfully chosen to add value to the participants' educational experience while ensuring that they do not exert undue influence over the decision to participate. This balance is crucial to maintaining the ethical integrity of the recruitment process and ensuring that participation is voluntary and unbiased.

1.5 Procedure

Participants are currently being sought for a usability study designed to evaluate the efficiency and user experience of a website across various educational levels, specifically targeting individuals in highschool, undergraduate, graduate and lecturers. We will be experimenting with two similar quiz platforms: QuizCreate, our in-house solution, and Quizlet, a popular external quiz website.

The study will be organized into a structured 8-week testing period, having each website spend 4 weeks each and participants from various educational levels, will spend 1 week on each section. Day 0 starts off with a detailed process

Independent Variables	Highschool Students	Undergraduate Students	Graduate Students	Professors/Teachers /Lecturers
QuizCreate	Group 1-1 Week 1	Group 2-1 Week 2	Group 3-1 Week 3	Group 3-1 Week 4
Quizlet	Group 1-2 Week 5	Group 2-2 Week 6	Group 3-2 Week 7	Group 3-2 Week 8

Fig. 2. Table of Participation Groups based on Independent Variable

for participants, providing a comprehensive overview of the study's objectives and tasks. This will collect pre-existing knowledge and participants' expectations and perceptions of the study. This will be formatted in a small survey and user agreement.

Day 1, participants are tasked with exploring various features of the website, encouraging interaction with different sections and functionalities. There will be 2 types of main data to be collected here, Duration spent on different pages or sections and the navigation path of sequence in which participants navigate through various features.

Day 2 shifts the focus to the creation of a quiz, aiming to evaluate the ease of use and efficiency of the quiz creation process. This would measure how quickly participants can create a quiz. Any challenges faced during the quiz creation process and which tools or options were utilized during quiz creation.

Day 3, participants engage in searching for quizzes to assess the effectiveness and speed of the search functionality. This would be very similar to day 2 however this would measure the search for quizzes section of the website.

Day 4 is designated as a free day, allowing users to explore the platform at their discretion. This day focuses on collecting data from the server end, identifying errors, and uncovering areas for improvement.

Day 5, participants take a quiz to assess the user experience. Following the quiz, they participate in a survey aimed at gathering subjective feedback on overall usability and satisfaction. This structured approach ensures a thorough examination of the website's usability and user experience between the 2 websites

Data compilation will involve the collection and analysis of both quantitative metrics such as task completion times and survey responses, as well as qualitative insights from participant interactions. Throughout the entire 4-week testing period, continuous server monitoring will be implemented to collect relevant server-side metrics, including response times, resource usage, and error rates. Additionally, specific data points will be gathered during the creation of quizzes

and search activities to identify potential usability problems. The time taken by participants to complete various tasks will be recorded, enabling us to pinpoint areas of improvement.

Following this, during weeks 2, 3, and 4, we will replicate the process and daily tasks for participants across all educational groups. Beginning in week 5, we anticipate that students who have had a sufficient 4 week break since week 1 will approach Quizlet with a refreshed perspective. We would hope that this result will be more accurate and efficient data collection.

1.6 Ethics

Prior to the commencement of the research sessions, ethical considerations will be carefully integrated into the study design. Participants will be required to sign a consent form, which serves as a formal agreement for their voluntary participation in data collection. This process involves thoroughly informing them about the study procedures, any potential risks, and specific aspects such as screen recording for data capture. It's imperative that participants understand their rights, including the option to withdraw from the study at any point without facing any penalties. During the research sessions, the utmost respect will be paid to the participants' time, ensuring that their experience is not overwhelming. The tasks assigned will be deliberately simple and essential to the study's objectives. For instance, when participants are asked to create a quiz, the instructions are kept flexible without demanding exhaustive details, thereby minimizing participant burden. A key focus will also be placed on maintaining participant privacy. In the case of surveys, measures will be taken to ensure anonymity, with data collection limited to essential information like date of birth for age categorization purposes. This approach has the power to minimize the personal data footprint and enhances participant comfort regarding privacy. After the completion of each session, stringent data security protocols are implemented. Participants will be well assured that the data collected is exclusively used for improving the system and will not be shared with any third-party organizations. This is beyond crucial for maintaining trust and transparency with the participants. Additionally, participants are debriefed about the data collection methods to ensure they are comfortable with how their information is being used. Finally, at the end of each session, participants will be thanked and congratulated for their valuable contribution and the time they dedicated to the study. This gesture of appreciation will not only be a courtesy but an acknowledgment of the critical role they play in the research process.

2 DATA ANALYSIS

2.1 successes

2.1.1 Descriptive Statistics. (Figure 3) For the success variable, I have chosen to calculate the mean, median and standard deviation. This is so that it was possible to see if there were any significant difference between "pm" and "pw". A first glance on the results tells us that there are no significant differences between the mean, median and standard deviation. The median is 2 and the average for both pw and pm are pretty close to that. The preliminary analysis of the data tells us that most users will succeed by the second attempt regardless of whether they written down or memorised their password.

Means

Report

successes

cond	Mean	N	Std. Deviation	Median
pm	1.67	27	.620	2.00
pw	1.93	28	.466	2.00
Total	1.80	55	.558	2.00

Fig. 3. Descriptive Data for the Success Variable

2.1.2 Data Visualization. The graph chosen was the clustered bar graph (Figure 4), as this will really highlight any differences between "pw" and "pm". This graph solidifies the results found in the earlier using the mean and median as it shows that majority of the users will succeed at the second attempt. Looking at the details though of the variables, "pw" had zero participants that had succeeded no times, and even had 3 successes. This study could benefit from using a higher sample size to see if different trends emerge, as the sample data right now has a small trend of "pw" being better, but not enough data points to be significant.

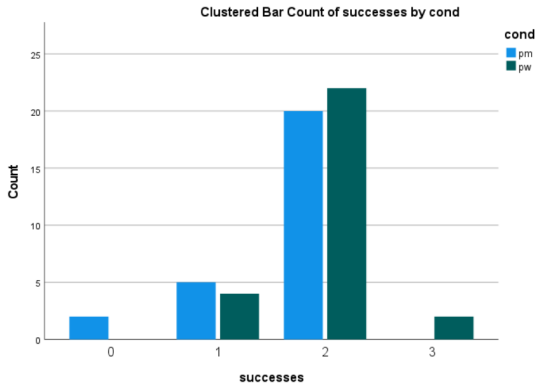

Fig. 4. Clustered Bar Graph showing the difference in the Success Variable

2.1.3 Inferential Statistics. The Mann-Whitney test was chosen to test the data for the success variable (Figure 5). This is because success is an ordinal data set, from 0 – 4, where 0 is the worst and 4 is the best case. This means the test used will be non-parametric and since the data is between subjects for 2 groups of participants, the Mann-Whitney test was chosen. The results are not significant (p = 0.123). We are looking at the two-tailed variant, as we do not know which way the test will go. Therefore, we will fail to reject the null hypothesis, which means that "pw" is not significantly different form "pm".

Mann-Whitney Test

Ranks

	cond	N	Mean Rank	Sum of Ranks
successes	pm	27	25.48	688.00
	pw	28	30.43	852.00
	Total	55		

Test Statistics[a]

	successes
Mann-Whitney U	310.000
Wilcoxon W	688.000
Z	-1.543
Asymp. Sig. (2-tailed)	.123

a. Grouping Variable: cond

Fig. 5. Mann-Whitney Test For the Success Variable

2.2 failures

2.2.1 Descriptive Statistics. For the number of failed logins over the study duration, the mean, the median, the mode, and the standard deviation are all going to be important descriptive statistics (Figure 6). These statistics will indicate which type of person is more likely to make a failed attempt of logging in, which quantity of failed attempts was the most common, and also how spread the distribution was.

Report

failures

cond	N	Mean	Median	Std. Deviation
pm	27	1.26	1.00	1.483
pw	28	.61	.00	2.315
Total	55	.93	.00	1.961

Fig. 6. Table of descriptive statistics, namely the mean, median, and standard deviation, for the number of failures between the two groups of participants

2.2.2 Data Visualization. To properly present the distribution of the data, we used a clustered bar graph (Figure 7), which compares the frequency of each number of failures between both groups. We can see a big difference in the results; "pw" participants committed much fewer failed attempts compared to "pm" participants. Also, the distribution of the results for "pw" participants is more right-skewed compared to the distribution of the results for "pm" participants.

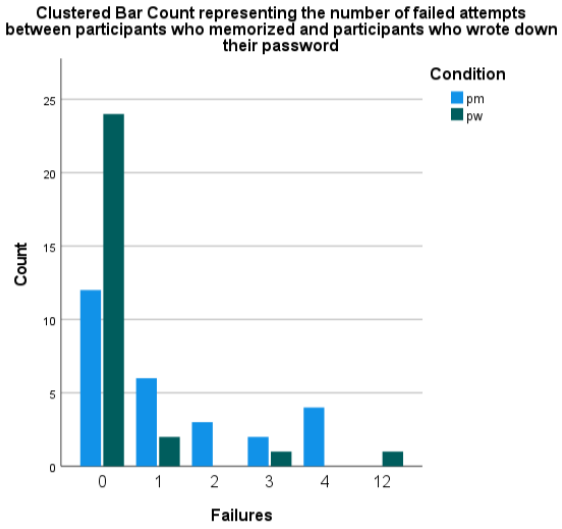

Fig. 7. Clustered Bar Count representing the number of failed attempts between participants who memorized and participants who wrote down their password

2.2.3 Inferential Statistics. For this dependent variable, we will be using the Mann-Whitney's U test (Figure 8). This is because the variable is ordinal; a lower number of failures is better than a higher number. The null hypothesis for this test is: There is no difference in terms of central tendency between the two groups in the population. So, if the means of the results are skewed enough, we can confirm that there is a significant effect.

Mann-Whitney Test

Ranks

	cond	N	Mean Rank	Sum of Ranks
failures	0	27	33.74	911.00
	1	28	22.46	629.00
	Total	55		

Test Statistics[a]

	failures
Mann-Whitney U	223.000
Wilcoxon W	629.000
Z	-3.084
Asymp. Sig. (2-tailed)	.002

a. Grouping Variable: cond

Fig. 8. Results after running Mann-Whitney's U test when comparing both groups' number of failed attempts

After running Mann-Whitney's U test, we can confirm that there is a significant difference (p = .002) between the central tendencies of both groups of participants (U = 223.0, Z = -3.084, p < 0.05, r = 0.42).

2.3 resets

2.3.1 Descriptive Statistics. We thought that the mean, median, mode and standard deviation are all appropriate data to represent the number of reset passwords (Figure 9). By having these 4 tests will give us enough information to see the average, most frequent and an appropriate spread of distribution.

Report

resets

cond	Mean	Median	Std. Deviation
pm	.30	.00	.609
pw	.04	.00	.189
Total	.16	.00	.462

Fig. 9. Results of the mean median and standard deviation of resets

2.3.2 Data Visualization. Used a cluster bar graph to represent the number of resets for both groups

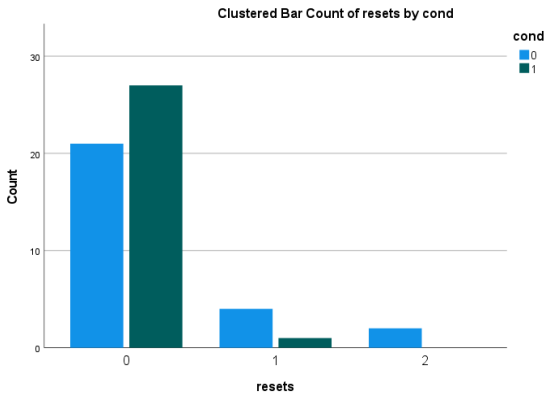

Fig. 10. Cluster bar graph comparing "pm" and "pw" with number of resets

2.3.3 Inferential Statistics. From the representative of the descriptive of resets, we can claim that the dataset is a non-parametric ratio since it is not normally distributed and the scale starts at a true zero. In this case, the 0 represents the number of password resets. We would be using a Mann-Whitney test to signify their distribution (Figure 11). According to the Mann-Whitney Test dataset for reset variables, there is minimal difference between the number of participants in “pm” compared to “pw”. (U = 306.0, Z = -2.081, p < 0.05).

Mann-Whitney Test

Ranks

	cond	N	Mean Rank	Sum of Ranks
resets	0	27	30.65	827.50
	1	28	25.45	712.50
	Total	55		

Test Statistics[a]

	resets
Mann-Whitney U	306.500
Wilcoxon W	712.500
Z	-2.081
Asymp. Sig. (2-tailed)	.037

a. Grouping Variable: cond

Fig. 11. Results after performing Mann Whitney Test between Two Sample t test for the unpaired groups' for Resets

2.4 sLoginTime

2.4.1 Descriptive Statistics. For the slogintime variable, I have chosen to calculate the mean, median and standard deviation (Figure 12). This is so that it was possible to see if there were any differences in the central tendencies between "pm" and "pw". The results were very different. "pm" had a very high standard deviation which meant that there was a higher spread of data. This variance means that "pm" has a less precise set of data, which shows that there is no consistency between getting a faster login speed for pm users.

Means

Report

slogintime

cond	Mean	N	Std. Deviation	Median
pm	9.0370	27	6.88215	7.0000
pw	9.8154	28	3.40427	9.5000
Total	9.4333	55	5.36224	8.0000

Fig. 12. Descriptive Statistics for the Slogin Variable

2.4.2 Data Visualization. The histogram (Figure 13) shows that the “pm” histogram is skewed right. Looking closer at the results shows us that this is due to outlier results and not an even spread of data. Removing the outlier results shows us that the distribution for pm and pw are both normal and basically the same.

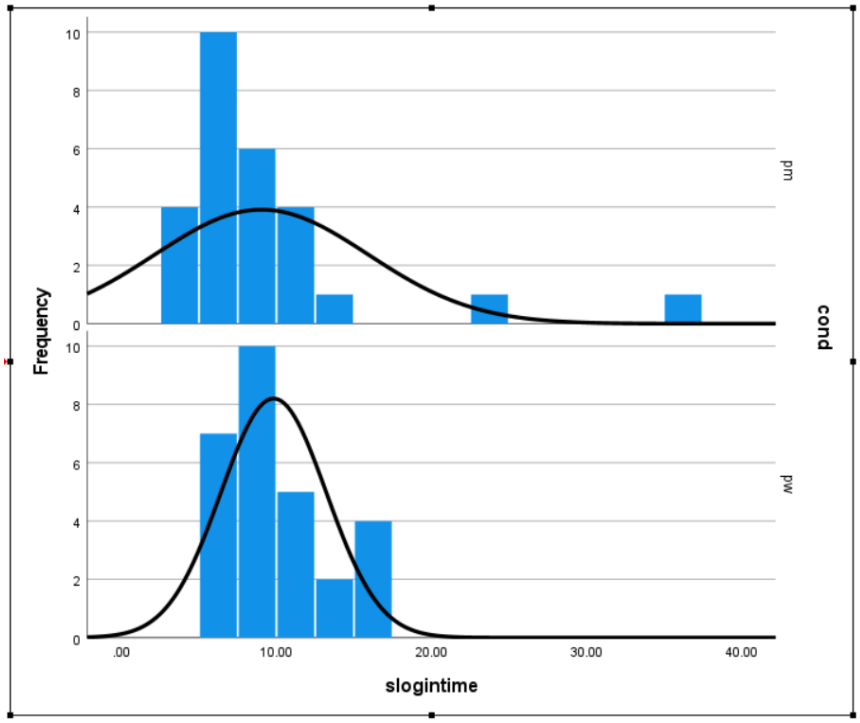

Fig. 13. Histogram showing the difference between "pm" and "pw" participants slogintime

2.4.3 Inferential Statistics. The unpaired t-test was chosen to test the data for slogintime (Figure 14). This is because slogintime is ratio data set measured in seconds for between subjects. Since we know that the standard deviation is high, we will go with the unequal variance. The results were not significant, therefore we will fail to reject the null hypothesis, which means that “pw” is not significantly different from “pm”. (p = 0.600)

T-Test

Group Statistics

	cond	N	Mean	Std. Deviation	Std. Error Mean
slogintime	pm	27	9.0370	6.88215	1.32447
	pw	28	9.8154	3.40427	.64335

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance		Mean	Std. Error	95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Difference	Difference	Lower	Upper
slogintime	Equal variances assumed	1.645	.205	-.535	53	.298	.595	-.77832	1.45599	-3.69866	2.14202
	Equal variances not assumed			-.529	37.696	.300	.600	-.77832	1.47245	-3.75993	2.20329

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	Upper
slogintime	Cohen's d	5.39806	-.144	-.673	.386
	Hedges' correction	5.47597	-.142	-.663	.380
	Glass's delta	3.40427	-.229	-.759	.306

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

Fig. 14. T-Test for the slogintime

2.5 maxMemTime

2.5.1 Descriptive Statistics. (Figure 15) For this dependent variable, a descriptive statistic that would be meaningful for us is the mean of all maximum memorizing times between both conditions. This will allow us to make a proper comparison between the average time of participants who have memorized or written down their password to log in properly since the creation of their new password. Furthermore, the median would be another meaningful descriptive statistic for this dependent variable. We can compare the results of the median for both groups with their respective means and check where the middle value is positioned and see whether the distribution is skewed in a certain direction or not.

maxmemtime

cond	N	Mean	Median
pm	27	169.0011	164.6100
pw	28	208.1464	184.8600
Total	55	188.9296	168.5400

Fig. 15. The descriptive statistics for the maximum memorizing time between the two groups of participants

2.5.2 Data Visualization. To visualize the data, we used a boxplot that compares the results for maxMemTime in both groups (Figure 16). The graph displayed information about the outliers, and where the upper and lower quartiles and extremes were situated relative to these results. Based on this representation, the boxplots seemed to be in similar positions with little change to the median.

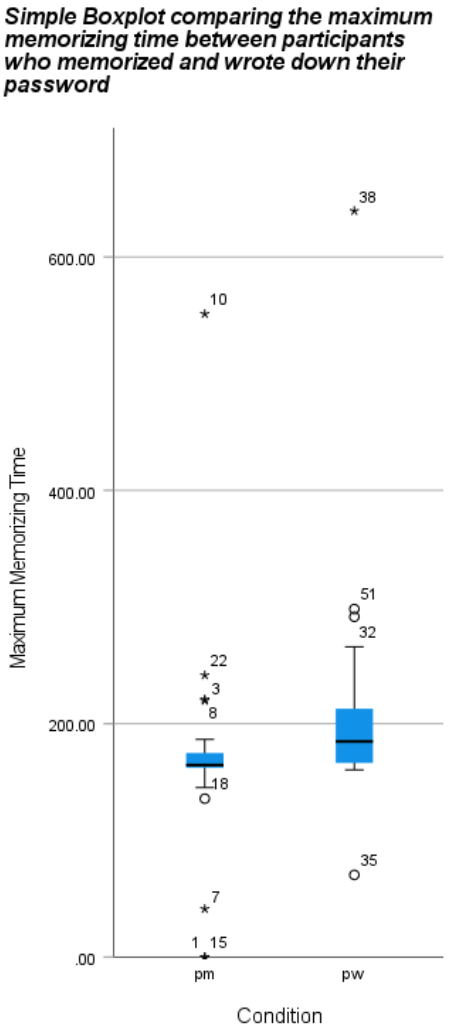

Fig. 16. Simple Boxplot comparing the maximum memorizing time between participants who memorized and participants who wrote down their password

2.5.3 Inferential Statistics. For this dependent variable, the unpaired t-test is the most appropriate test to compare the performance between the two groups of participants because both groups are two different groups of people. In other words, each subject was measured once. However, despite the variance values being very similar yet not identical, we must also perform Welch's t-test to accommodate the difference. Our null hypothesis is there is no significant difference in the means between the two groups. Here are the results of the t-test, calculated using R:

With a Welch's t-test (Figure 17), we did not find a significant effect for password memorization ($p = 0.1334$). Therefore, the null hypothesis cannot be rejected. In other words, we cannot confirm that there is a significant difference in the means between the two groups ($t(53) = 1.52$, $p > 0.05$).

```
        Welch Two Sample t-test

data:  data[data["group"] == 0, 2] and data[data["group"] == 1, 2]
t = -1.5242, df = 52.968, p-value = 0.1334
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -90.65728  12.36665
sample estimates:
mean of x mean of y
 169.0011  208.1464
```

Fig. 17. Results after performing Welch's Two Sample t-test for the unpaired groups' maximum memorizing time

2.6 totalLastingSuccess

2.6.1 Descriptive Statistics. For totalLastingSuccess we chose the mean and median as the most efficient way to represent this dataset (Figure 18). Since there are only 2 numbers of success/failure we would want to find which of the numbers is more frequently seen and how often.

Report

totalLastingSuccess

cond	N	Mean	Median
pm	27	.78	1.00
pw	28	.86	1.00
Total	55	.82	1.00

Fig. 18. Results after calculation of mean and median

2.6.2 Data Visualization. We used a cluster bar graph to show the proportion of 0s and 1s in the dataset because it is more effective at emphasizing the proportion of each category relative to the whole (Figure 19).

2.6.3 Inferential Statistics. Nominal data represents categories without any inherent order or hierarchy. In this particular data of 0s and 1s which represent success or failure. Therefore, by definition of a nominal set this would be the most efficient way to show this type of statistical data (Mann-Whitney test). Using the data of the Mann-Whitney test (Figure 20), we can say that the null hypothesis is rejected representing a minimal difference between the two groups (U = 348.0, Z = -.756, p = 0.45, r = 0.11).

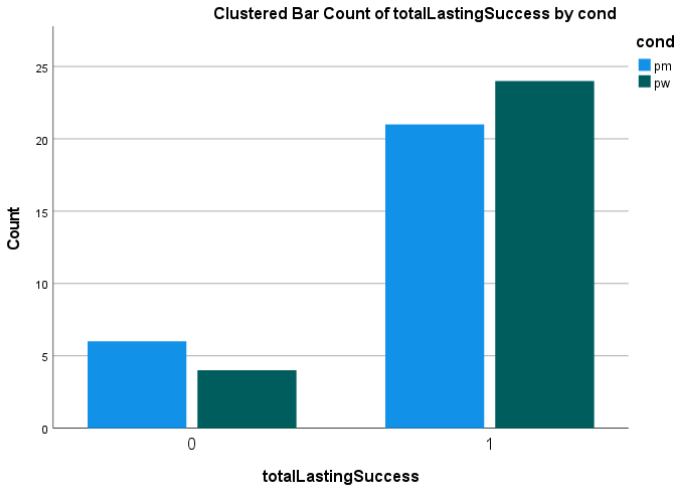

Fig. 19. Graph of totalLastingSuccess of both data groups

Ranks

	cond	N	Mean Rank	Sum of Ranks
totalLastingSuccess	0	27	26.89	726.00
	1	28	29.07	814.00
	Total	55		

Test Statistics[a]

	totalLastingSuccess
Mann-Whitney U	348.000
Wilcoxon W	726.000
Z	-.756
Asymp. Sig. (2-tailed)	.450

a. Grouping Variable: cond

Fig. 20. Results after performing Mann Whitney Test between Two Sample t test for the unpaired groups' for totalLastingSuccess

2.7 avgLeadingFailures

2.7.1 Descriptive Statistics. For the avgleadingFailures variable, we have chosen to calculate the mean, median and standard deviation (Figure 21). This is so that it was possible to see if there were any differences in the central tendencies between "pm" and "pw". The results were very different. "pw" had a very high standard deviation which meant that

there was a higher spread of data. This variance means that “pw” has a less precise set of data, which shows that there is no consistency between getting a faster login speed for pm users.

Means

Report

avgLeadingFailures

cond	Mean	N	Std. Deviation	Median
pm	.500	27	.7206	.000
pw	.286	28	1.1420	.000
Total	.391	55	.9559	.000

Fig. 21. Descriptive Statistics for avgLeadingFailures

2.7.2 Data Visualization. The histogram (Figure 22) shows that the “pw” histogram is skewed right. Looking closer at the results shows us that this is due to outlier results and not an even spread of data. Removing the outlier results shows us that the distribution for "pm" and "pw" are both normal, but that "pw" would have a way lower standard deviation as the results are basically all in one bar. This means that “pw” has a lower number of average leading failures by a higher margin compared to pm.

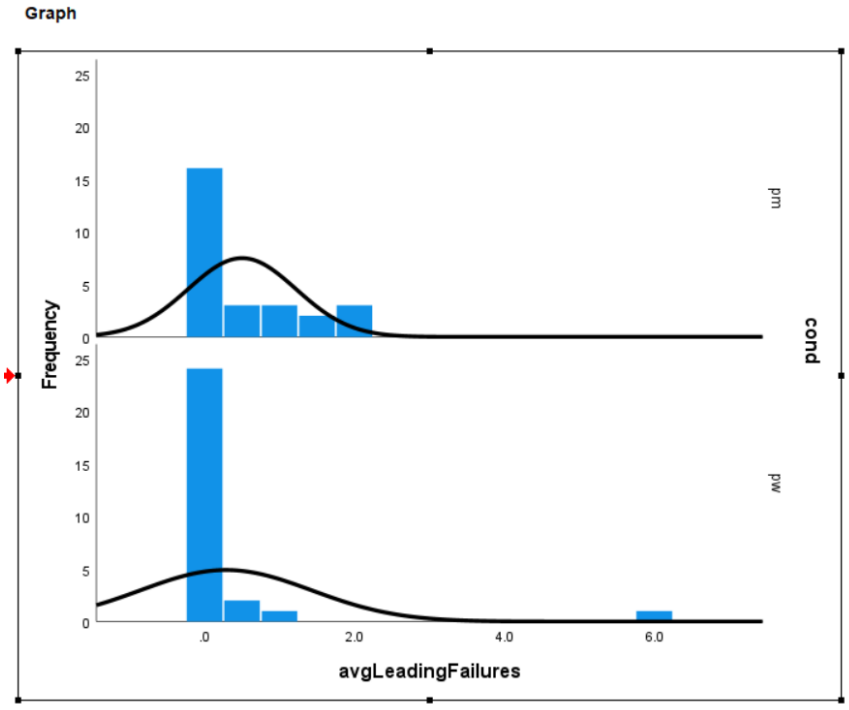

Fig. 22. Histogram between "pw" and "pm" participants in terms of the avgLeadingFailures

2.7.3 Inferential Statistics. The Mann-Whitney test was chosen to test the data for the avgleadingFailures variable (Figure 23). This is because avgleadingFailures is a ratio, but the data set is not distributed normally. We are looking at the two-tailed variant, as we do not know which way the test will go. The (p = 0.026) which is lower than (p = 0.05) which means the results are significant. Therefore, we reject the null hypothesis, which means that "pw" is significantly different form "pm". (p = 0.026).

Mann-Whitney Test

Ranks

	cond	N	Mean Rank	Sum of Ranks
avgLeadingFailures	pm	27	31.83	859.50
	pw	28	24.30	680.50
	Total	55		

Test Statistics[a]

	avgLeadingFailures
Mann-Whitney U	274.500
Wilcoxon W	680.500
Z	-2.223
Asymp. Sig. (2-tailed)	.026

a. Grouping Variable: cond

Fig. 23. Results of the Mann-Whitney Test for avgLeadingFailures

2.8 Discussion

2.8.1 Failures. There is a minimal effect on the memorability and usability of passwords. One of the only significant changes is the number of failures, where participants who wrote down the passwords failed less compared to the participants who memorized their password. According to the graph, the most common number of failed attempts for both "pm" and "pw" groups is zero. Moreover, observing the mean tells us that the average number of failed attempts is twice greater for memorized passwords compared to written passwords. The standard deviations are also different; the distribution for failures are much more right-skewed for written passwords compared to memorized passwords, meaning there is generally a lower failure rate when attempting to login when people write down their passwords.

2.8.2 avgLeadingFailures. Furthermore, the avgLeadingFailures before successfully logging in is important to note as well; users who wrote down their passwords have a low number of failed attempts when logging in compared to those who have not. While both "pw" and "pm" lie in the same area in the histogram, the distribution is very spread out for "pm", while most of "pw" lies in 1 bar. This means that avgLeadingFailures shows a significantly lower for "pw" than "pm" participants.

2.8.3 Resets. There is not a big significant difference between the number of users resetting their passwords by the means/median/mode. Based on the graph, more people are likely to not reset their passwords and there are very few people who reset their password more than once over the duration of the study period. This could be because users don't seem to need to reset their password because of lack of security concerns and the unlikelihood of their accounts being breached.

2.8.4 totalLastingSuccess. There is little difference between the "pm" and "pw". According to the graph, there seem to be more users that kept their own password for the duration of the study period and a small number of users changing their passwords. This could be because users don't seem to need to change their password because, if users find their passwords memorable or convenient to use, they might be less motivated to change.

2.8.5 Successes. There was no significant difference between "pm" and "pw" participants in the descriptive, inferential, or graph for the success variable. The clustered bar graphs show that most of the participants lie on the 2 successes out of 4. This lack of difference is strange especially due to the fact that the number of failures for "pw" is significantly lower than "pm". This can be accounted for by the users not using the written passwords initially, meaning they don't succeed on the first try but refer to the written password in the second attempt which averages out the successes and lowers the failures.

2.8.6 Slogintime. There was no significant difference between "pm" and "pw" participants in the descriptive or inferential data for the "slogintime" variable. The histogram shows that "pm" is skewed right which might mean that "pm" is faster at logging in. But when the histograms are compared, you can see that both graphs overlap and have the curve at the same exact position. This is because the skewed graph is caused by an outlier. The reason that there are no significant differences is because participants who memorized the password had a higher margin of error and the written password participants took the time to look at the password. This has resulted in approximately the same amount of time taken to login successfully between both participant groups.

2.8.7 MaxMemTime. Comparing both results, we can clearly see a difference in average time between both groups of participants; to correctly log in to their account since the creation of their new password, those who have memorized

their password generally take 40 seconds less than those who have written down their password. Yet, this difference was deemed insignificant after performing the statistical test. This might be because of each groups' outliers; some participants who have memorized their passwords may have used a browser extension to help memorize their password without writing it down. In conclusion, if users write down their passwords instead of memorizing, the only difference it makes is the lower frequency of errors.

3 DISTRIBUTED SUMMARY

3.1 Garrison Distribution

(1) User Study Design
    (a) Procedure
    (b) Study Design
(2) Data Analysis
    (a) Resets
    (b) totalLastingSuccess
    (c) Discussion (For all dependent variables above)

3.2 Tooba Distribution

(1) User Study Design
    (a) Analysis Plan
(2) Data Analysis
    (a) Success
    (b) SLoginTime
    (c) avgLeadingFailures
    (d) Discussion (For all dependent variables above)

3.3 Merraj Distribution

(1) User Study Design
    (a) Study Design
    (b) Data Collection
(2) Data Analysis
    (a) Failures
    (b) MaxMemTime
    (c) Discussion (For all dependent variables above)

3.4 Arda Distribution

(1) User Study Design
    (a) Ethics
    (b) Participants

4 REFERENCES

[1] Nielsen Norman Group. 2016. *The Distribution of Users' Computer Skills: Worse Than You Think*. Retrieved from https://www.nngroup.com/articles/computer-skill-levels/

[2] Nielsen Normal Group. 2008. *Middle-Aged Users' Declining Web Performance*. Retrieved from https://www.nngroup.com/articles/middle-aged-web-users/