Hi-Fi Prototype Usability Study for Calendar Application with Timed Event and Statistics Functionalities

George Covalciuc (gco203), Yasin Karyağdi (yki430),
Laura Stampf (lsf270), Tomáš Schuster (tsc211)

Abstract

This study examines the impact of user interface design on the usability of a calendar application, with a specific focus on timed events and statistics functionality. The user study uses a between-subjects design, dividing participants into two distinct groups based on their prior experience with the application's user interface (UI) of a high-fidelity (hi-fi) prototype - "Some" or "None." Each participant is tasked with completing four tasks. The study evaluated how user experience levels with the hi-fi prototype influence overall performance (task completion time and number of clicks). Our findings indicate a significant difference in task completion times between groups, with experienced users completing tasks in an average of 37.54 seconds, compared to 61.185 seconds for inexperienced users—a 62.98% increase. Additionally, for the number of interactions we also found that the experienced group performed better, with a 40% difference (20 clicks for experienced users versus 28 for inexperienced ones). This suggests that user performance is always higher once they are accustomed to the UI. Statistical analysis, using the Mann-Whitney-Wilcoxon test, confirmed the significance of the difference in task completion times ($p = 0.004434$) but not in the number of clicks ($p = 0.7307$). Findings reveal that while experience impacts time taken to perform tasks initially, it seems not to significantly affect the amount of clicks needed to navigate the interface. However, the study encountered several limitations, like having a small, non-diverse sample of candidates and incomplete app features affecting the results.

Keywords

Between-Subjects Study, Calendar Applications, Hi-Fi Prototype, Human Computer Interaction, Human Factors, Interaction Design, Performance Metrics, Statistical Analysis, User Interface (UI), User Feedback, Usability

1 Introduction

Despite the wide array of calendar applications available today, users frequently encounter difficulties in effectively managing timed events, often struggling with unintuitive interfaces that complicate rather than help in their daily time management tasks. Recognizing these difficulties, we developed an application designed to address these challenges.

1.1 Use Case Description

Our application is a calendar that has the added functionality of tracking time and displaying statistics to its users. The goal of the system is to provide functionality for general time and task management. The application is targeted towards both casual users who want to use basic features, as well as experienced users who use advanced features such as graph customization. To fulfill this goal, the calendar allows basic features such as adding, removing, and editing events. Additionally, it enables time tracking by featuring timed events (in the form of a timer or stopwatch) that can be linked to specified activities. To give users further insight into their time usage, the application also provides a dynamic statistical overview to help optimize one's schedule.

The application features two main views. In the calendar view, the user can interact with any event or with the timed events. One can easily add events, modify them, and start a time-tracked session, all within this panel. In the second view, the statistical display, the user can filter the data that is shown. The statistical overview provides all the tools the user needs to get an accurate assessment of how they spend their time, whether it is filtering by event category or modifying different graphs and charts.

Note that we do not target one specific audience; all kinds of users may use the application. The users themselves decide how much functionality they want to use. One such example would be two students using the application. One student can decide to only use the application for adding, removing, and editing events, totally ignoring the statistical overview functionality. While the other student can choose to fully make use of the statistical overview functionality and to always use the stopwatch feature while studying.

1.2 Design Issues Identified from Pilot Study

While the application's diverse features, as described in our use case, aim to cater to a wide range of users, they also introduce a set of design challenges that need careful consideration. In our preliminary efforts to refine the app's usability, we conducted a pilot study, seeking direct feedback from a small user group. This initial investigation brought several critical interface design issues to light:

- The interface design for the statistical view often becomes cluttered, potentially overwhelming users unfamiliar with complex data analytics.
- Simple actions, like changing visualizations in the statistical view, can be daunting for casual users, leading to decreased engagement.

- A minimalist feature set, while appealing for simplicity, could strip away the customization options critical for power users.

Additionally, feedback from the pilot study highlighted specific areas for improvement:

- Confusion around the bookmarks section suggested a need for clearer functionality and renaming to 'presets' for better user comprehension.
- The absence of a current time indicator in the calendar view was identified as a major oversight, making it necessary to add a visual cue, such as a red line, to enhance the sense of "now."
- The statistical overview's clarity was called into question, with users asking for percentages on the pie chart to visualize better data interpretation.

1.3 Hi-Fi Prototype Description

Guided by the insights gained from our pilot study, we started on the development of our hi-fi prototype, addressing some of the highlighted design issues to craft a more intuitive and user-friendly application

Our hi-fi prototype has been designed for this paper's usability study. Not all functionality and screens presented in our lo-fi design are supported. For example, the settings page and event adding, modification, and deletion have not been implemented in this hi-fi prototype. Nevertheless, the addition of these features may be useful for future experiments. The results of the pilot study guided several changes to the design choices in our hi-fi prototype. The following highlights the main design choices:

Timer/Stopwatch Toggle The pilot study aimed to clarify the difference between setting a timer and a stopwatch. From the experimental results we gathered, it was clear that our initial lo-fi prototype was flawed - participants had trouble distinguishing between both types of timed events. Nevertheless, the second iteration of our lo-fi prototype received positive feedback. Thus, we chose to use this alternative UI as a starting point for an improved UI. From the verbal feedback received from participants, we identified that it was not necessary to display the time slider when the Timer is toggled, so this was removed entirely from our hi-fi design. It should be noted that, though the Timer/Stopwatch functionality's appearance is implemented, it is not "responsive"; the calendar and statistical overview do not dynamically update upon saving a timed event - this is not needed, it supports enough functionality for us to reasonably perform our user study.

Bookmarks/Presets From the feedback received during the pilot study, we learned that the Bookmarks feature was unclear. Participants reported that the Bookmarks icon was misleading and the general functionality was ambiguous. To address these concerns in our updated hi-fi design, we had to remind ourselves of the purpose of the Bookmarks feature: to help users initialize commonly reused timed events more quickly, without needing to enter all fields manually. From this reflection, we recognized that the naming of this feature could be enhanced to add clarity. Instead of "Bookmarks", we chose to change the naming to "Presets". Additionally, we decided to change the icon to a star, as recommended by one of the participants of the pilot study. To further enhance the clarity of this "Presets" functionality, we aim to add a short description at the top of the Presets window that is triggered by clicking the star icon. Due to time constraints, the Presets functionality is not yet implemented fully, the only visible change has been the icon change. Nevertheless, we intend for the general appearance to remain similar to our previous prototype, with changes from the name "Bookmarks" to "Presets" and with the addition of a short description at the top of the window.

Statistical Overview Data Visualization One participant of the pilot study mentioned that aside from the visual indication of time distribution in the pie chart, there should be numerical labels for more precise time estimates. Therefore, to enhance the clarity of the pie chart proportions, our hi-fi prototype incorporates two numbers per slice: an absolute measure of the number of hours a certain activity took and its relative proportion of the total tracked time. Additionally, to give precise time measures of each bar in the bar chart, we added the additional feature of clicking a specified bar to reveal precise hours. This added user interaction on the bar chart enhances the user experience by providing easy access to these numerical features. The UI of the Statistical Overview closely resembles that presented in the low-fi prototype with the inclusion of explicit data labeling in the pie chart and bar chart.

1.4 Motivation

The purpose of the user study is to identify issues that users with different experience levels may face regarding our hi-fi prototype and to mitigate these through design changes in future iterations. Additionally, we want to gain insight into the ease of navigability of our application. A key component of our design choices has been simplicity. We want to provide users with a seamless experience with our application: icons should be informative enough and act as a complete replacement for textual information, and the use of buttons, toggles, filters, and other features should be appropriate for easy use by even inexperienced users. We aim for our application to be used effectively regardless of the experience level of our users. Our study will focus on measuring how the independent variable, the experience level of participants, impacts their performance on a given task in regards to the dependent variables of time taken and number of clicks required. Through this user study, we strive to gain both empirical but also verbal feedback to ensure that ease of use is fulfilled. If we find that

our current prototype does not meet these requirements by for instance seeing a great performance difference between both experience-level groups, we will adapt our design to satisfy our user's needs.

This study examines how users with different levels of exposure to our UI, those with no prior exposure and those with some familiarity, interact with our high-fidelity prototype, identifying usability issues with the intent of enhancing its design to improve user interaction.

2 Method

2.1 Participants

Participants of our user study will be chosen on a convenience-basis, due to the project's limited scope. For this reason, all participants will be limited to VU Computer Science students between the ages of 18-25. Nevertheless, though this narrow participant selection only highlights a subgroup of the application's user base, this chosen demographic also presents some benefits. Computer Science students are likely more tech-savvy than the average person. Exposure to a larger variety of technologies has the potential to help people identify personal preferences more clearly. In this sense, participants of our study may have more experience with similar calendar-like applications which can be immensely helpful in noticing potential improvements in our application.

Due to time restrictions, we aim to conduct the between-subjects study on 8 participants in total. We are measuring two different conditions: having no or some experience with the application, with each condition being tested on 4 participants.

2.2 Apparatus

A detailed list of all required materials for the user study is given below:

Participant laptop The experiment will be performed on one of the researcher's laptops, an Apple Mac Book Pro. All participants will be provided with the same device, without the need for an external mouse or keyboard; the laptop's native keyboard and trackpad will be used as input devices.

Researcher laptop The researcher conducting the user experiment will take note of observations, and feedback, and note down the numerical values of the dependent variables on a laptop throughout the experiment.

Hi-fi prototype The hi-fi prototype will be loaded on the participant's laptop. Each participant is presented with the same prototype.

Consent form Each participant will be asked to sign a consent form before proceeding with the experiment to ensure that they are aware of any ethical considerations to be made. The consent form provided to participants can be found at the following link. This document is provided to participants as a print out for participants to read through and sign. The consent form used in this user study is based on a template provided by the University of Twente [4]. For this, we will provide participants with a pen.

Click Counter application The Mac Mouse Click Counter application (on the participant's laptop) will be used to measure the number of clicks a participant uses to complete a given task. The researcher will note down the value measured after each task is complete.

Stopwatch The participant's task completion time will be recorded by the researcher using a stopwatch accessed from the researcher's computer.

2.3 Procedure

Our user study uses a **between subjects** design with two groups of participants of different experience levels, or conditions ("Some" or "None"). Participants of both conditions will be asked to perform the same four tasks. It is important to note that all tasks should be initialized from the same page across participants to ensure accurate results - every task should be started from the Calendar page. This further enables the measurement of navigability between both main pages of the application (Calendar and Statistical Overview). The following lists the four tasks that all participants will perform in the given order.

1. Start a stopwatch for the event "HCI Project" and label it as "Work"
2. Start a 2-hour timer for the event "Movie" and label it as "Family"
3. Report the number of hours spent on work this month
4. Report the days that you spent the most and the least time learning this week

The first two tasks are focused on assessing the usability of initializing a timed event (stopwatch or timer). Through these tasks, we aim to collect data regarding the clarity of the icons that are interacted with and of the use of a toggle to enable the separation of both timed events within the same window. This will be done by a performance comparison of both experience-level groups (in terms of the dependent variables) as well as through the collection of informal verbal feedback from participants.

The last two tasks require users to interact with elements of the Statistical Overview page to retrieve data about their time usage. Task 3 is focused on the pie chart's functionality and requires users to interact with the drop-down menu to correctly view the monthly overview of time distribution. Task 4 requires interaction with the week-based bar chart and optional filtering for specific activities. Both of these tasks will provide valuable feedback regarding the types of statistical visualizations used as well as adjacent functionality such as filtering or item placement.

2.4 Design

2.4.1 Variables

We have identified the following variables regarding our user study.

- **Independent variables**: The independent variable is the experience of the user with our UI with two conditions: either "None" or "Some". We will simulate this by giving a tutorial to half of the participants while giving no guidance to the other half. The tutorial will consist of an explanation, given by a researcher, on each of the application's pages and the functionalities corresponding to each icon.
- **Dependent Variables**: The dependent variables are the total time taken (in seconds) and the total clicks performed to complete a given task; both depend on the experience that the user has with the interface.

  Time taken This will be measured manually using a stopwatch. The stopwatch will be started as soon as the participant declares "go" and stopped as soon as the task is completed. After each task, the researcher will take note of the measured time.

  Number of clicks This will be measured by the use of the Mac Mouse Click Counter application. The counter will be initialized to 0 after the task is read out. As soon as the participant completes the task, the researcher will take note of the number of clicks before moving on.

- **Control variables**: The interface is an important control variable; each participant will perform identical tasks on the same interface of the hi-fi prototype with the same initial starting page. Another important control variable is the device used; all participants will use the same model of computer (Apple Mac Book Pro) with its built-in input methods (trackpad and keyboard). Additionally, the consent form that participants are provided will be consistent across all participants.
- **Uncontrolled variables**: Important uncontrolled variables that are likely to affect the study results are the technical experience that a participant has with similar applications in general and the participant's current state. We expect that participants with more tech-related experience will outperform their counterparts regardless of experience with our specific application as they may be more familiar with how to navigate the interface. The participant's state in terms of mood and attention is also something that likely will impact their performance regarding the dependent variables. We will attempt to somewhat reduce the impact of the latter factor by ensuring the experiment is conducted in a quiet environment to minimize distractions and present a positive and energetic attitude when presenting the tasks to engage the participant.

2.4.2 Data Collection & Storage

Our user study involves the collection and storage of the following items:

- **Consent forms**: Before initializing the actual experiment, we ask participants to fill in a consent form. Personal data collected in these include the participant's name and signature. Each participant is given the option to keep a copy of the consent form if desired. Additional information in this form includes the date and several consented items that have been check-boxed by the participant. As mentioned in the consent form itself, participant consent forms will be stored for 1 month electronically after which they will be discarded. The physical copies of the consent forms are discarded after scans have been made.
- **Participant performance results**: Throughout the experiment, the researcher will note down the performance metrics for the dependent variables (number of time taken and clicks) of each unique task presented to the participant. This numerical data is gathered electronically on the researcher's computer. Additionally, the results are anonymous; participant names are excluded and rather referred to by an ambiguous number (i.e. "Participant 1" instead of "Laura"). The raw performance data gathered will be stored electronically in a file on the researcher's computer for an indefinite amount of time.
- **Research notes on participant feedback and performance**: Aside from collecting numerical values of (anonymous) participant performance, the researcher also takes notes about how participants accomplished different tasks such as information about what buttons were (mis-)pressed. By noting down such intricacies, we can ask specific feedback points from participants customized to their interaction with the application. For example, participants who click the wrong icon to initialize a timed event can be asked "How can we make icons clearer?". The researcher will additionally ask participants for more holistic feedback regarding aspects of the hi-fi prototype including usability, navigability, and aesthetics. Similar to the storage of performance results, researchers will save these notes anonymously and electronically.

3 Results & Discussion

This section is separated into four subsections:

1. Raw Results
2. Averaged Results
3. Statistical Testing
4. Feedback Results

Subsections 3.1-3.3 discuss the different steps that were taken to analyze the numerical data gathered from the user study, especially concerning the dependent variables that were measured. Subsection 3.4 discusses results gathered from the verbal feedback received from participants throughout the experiment and based on common performance patterns.

3.1 Raw Results

The results from the user experiment are displayed in two separate Figures (1 and 2), each corresponding to the relevant dependent variable (task completion time or number of clicks respectively). Both experience conditions are identifiable by color; orange identifies participants with no experience and blue for some experience.

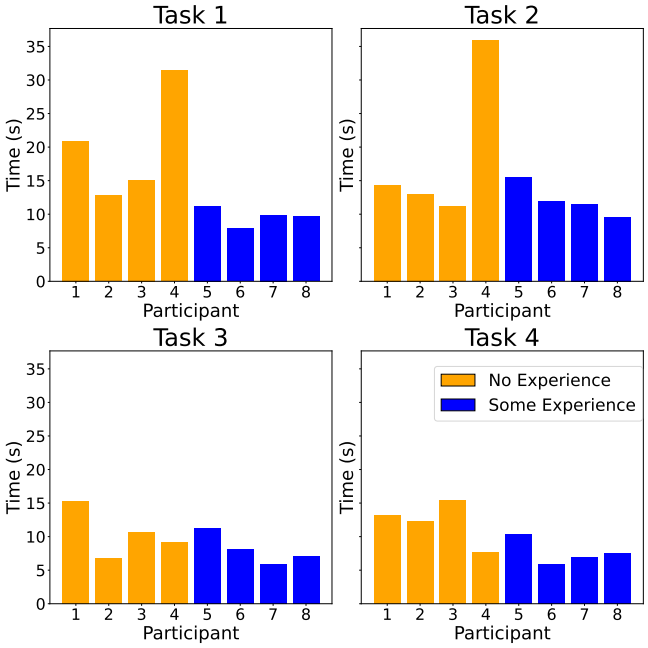

Figure 1: Time Taken Per Participant by Task

For both histograms plotting the gathered data in Figures 1 and 2, there is clear evidence of an outlier for Tasks 1 and 2, namely Participant 4. Both task performance measurements indicate that the participant performed significantly worse than other participants of the same condition. This distinction is due to a combination of technical issues encountered and the limited functionality inherent to the hi-fi prototype.

When completing Task 1, Participant 4 faced issues with the trackpad. Unlike other trackpads, Mac Book trackpads (also known as "Force Touch" trackpads) respond to different levels of pressure: exerting more force is registered as a "Force click" commonly used for lookups [2]. Due to the participant's inexperience with Mac Books, they accidentally used this feature to trigger the dictionary application's opening. This unexpected event led to additional time and clicks to resolve the issue and bring the participant back to the desired prototype window used for testing. This event highlights the importance of planning carefully for these kinds of unforeseen events and raises the question: what is the ideal control environment to prevent such mishaps in the

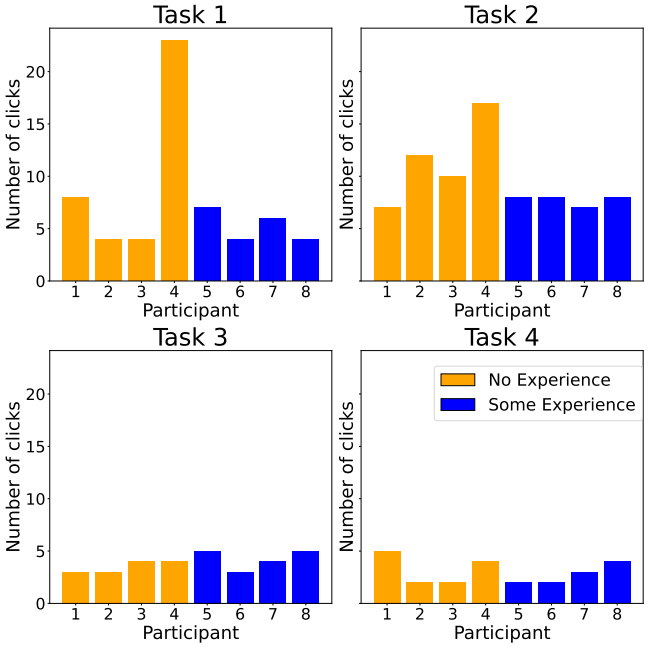

Figure 2: Number of Clicks Per Participant by Task

future? Does the use of a mouse instead of a trackpad reduce the number of input-related issues in experiments?

Participant 4's poor performance in Task 2 is less attributed to the input device used and more indicative of ways that different people interact with functionalities. The participant repeatedly clicked a button that did not respond due to not having been implemented. This phenomenon is related to the individual's experience with technology and how they respond to unresponsive functionality; other participants who clicked an unimplemented button quickly understood the lack of its functionality and hence their performance was less impacted by an initial error. Moreover, this occurrence highlights the importance of following Shneiderman's 8 Golden Rules for UI Design (from [3] as mentioned in the Lectures), especially regarding rule 3 of offering informative feedback to users and rule 7 of letting users feel in control. With this in mind, it is a good idea for any future studies to provide feedback for unimplemented buttons, such as the triggering of a window with text specifying "This button is not implemented. Please try other options".

3.2 Averaged Results

Aside from the clear outlier from Figures 1 and 2, there is little insight to be made about the general performance differences between conditions. To give a better overview of potential differences and to more easily identify potential trends, we plot the averages of both performance metrics in Figures 3 and 4.

From Figures 3 and 4 we see the average performance metrics demonstrate that participants with some experience outperformed the latter group on all tasks except for the number of clicks used on Task 3. However, it is important to consider the variability in our data, represented by the error bars. Again, we see the significant impact that the outlier (Participant 4) has on the re-

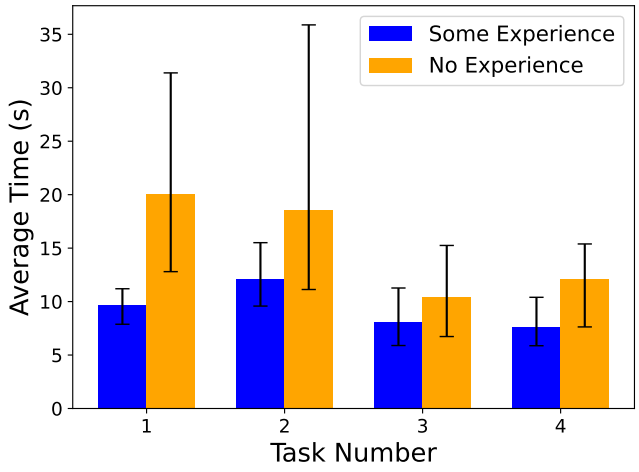

Figure 3: Average Time Taken Per Task by Experience Level

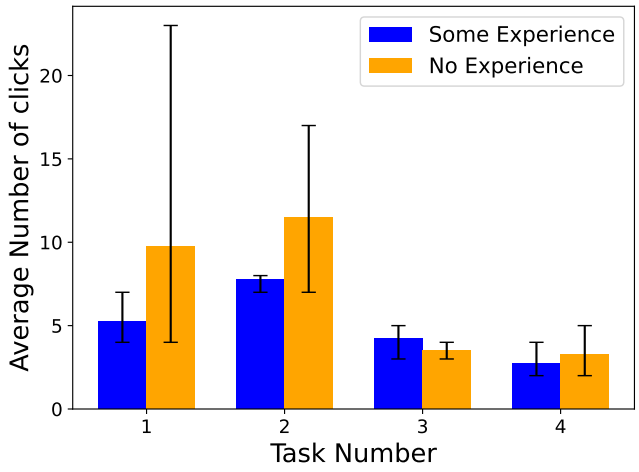

Figure 4: Average Number of Clicks Per Task by Experience Level

sults of tasks 1 and 2 in both figures, considering the height of the error bars for participants with no experience. The overlapping error bars between both groups per task make it clear that more rigorous variability testing is required.

Sum of means	**Time (s)**	**Number of clicks**
Experience	37.54	20
No Experience	61.185	28
Relative difference	62.98%	40%

Table 1: Results after summing the mean for each task. The relative difference has been calculated for both dependent variables

We summed the means of all the tasks to observe the overall performance difference between the two groups. We did so for both dependent variables and calculated the relative performance difference between the two groups (see Table 1). We can observe that the experienced users performed better both in terms of time taken and number of clicks performed. This aligns with our previous observation that for most tasks the experienced group performed better, though the addition of these comparative metrics allows us to quantify the individual difference between both conditions. We find that the participants without experience took about 62.98 % longer to complete all the tasks and they did so while using 40% more clicks. Nevertheless, it is unclear if these results are statistically significant and therefore they will need to be analyzed further using statistical testing.

3.3 Statistical Testing

We want to perform statistical tests on each dependent variable to quantify whether the variation between conditions is statistically significant. Before conducting any statistical tests, we frame our null hypothesis (H_0) and alternative hypothesis (H_a) as follows:

H_0: there is no difference in the distributions of both groups
H_a: there is a difference in the distributions of both groups

To test these hypotheses, we use a **Mann-Whitney-Wilcoxon** Test; this is suitable for our between-subjects study design with only two conditions (as mentioned in the book). The non-parametric test has been chosen as we do not assume normality in our data. We make use of the `wilcox.test()` function provided by R to conduct these tests.

First, we perform a test for the total performance difference between conditions on both dependent variables (see Table 2). The test results give us insight into the statistical significance between both conditions based on the dependent variable. We find that for the Number of clicks, there is no significant difference ($p > 0.05$), suggesting that the experience level does not significantly affect the total number of clicks needed to perform all tasks; there is not enough evidence to reject H_0. However, for the Time(s) we find that $p = 0.004434$. This result indicates that there is a statistically significant difference between the two groups ($p < 0.05$), suggesting that the experience level affects the time taken to complete all the tasks. Thus, in this case, H_0 is rejected in favor of H_a. To gain further insights into why this significance might occur, we performed an additional eight tests to look at the significance of task-dependent performance.

Dependent Variable	**W**	**p-value**
Time (s)	204	0.004434
Number of clicks	137.5	0.7307

Table 2: Mann-Whitney-Wilcoxon Test Results by Experience level

We will perform tests by considering the performance for each task individually. With two conditions and four unique tasks, we must conduct eight tests in total. The resulting W test statistic and p-value for each combination of task and dependent variable are provided in Tables 3 and 4.

The test results from both tables give us insight into the statistical significance between both conditions

Dependent variable	Task 1		Task 2	
	W	p-value	W	p-value
Time (s)	16	0.03038	11	0.4705
Number of clicks	10	0.6446	12.5	0.2338

Table 3: Mann-Whitney-Wilcoxon Test Results by Experience level for Tasks 1 and 2

Dependent variable	Task 3		Task 4	
	W	p-value	W	p-value
Time (s)	11	0.4705	15	0.0606
Number of clicks	4	0.2849	9.5	0.7568

Table 4: Mann-Whitney-Wilcoxon Test Results by Experience level for Tasks 3 and 4

based on the dependent variable and task being measured. Most tests do not show a statistically significant difference ($p > 0.05$), suggesting that the experience level does not significantly affect the task completion time or number of clicks; there is not enough evidence to reject H_0. There is however one exception: $p = 0.03038$ for the test results of Task 1 with dependent variable "Time (s)". These results indicate that there is a statistically significant difference between the two groups ($p < 0.05$), suggesting that the experience level affects the time taken to complete Task 1. Thus, in this case H_0 is rejected in favor of H_a.

Considering what may have led to a statistically significant test result for the time of Task 1 raises the question of the test data's validity. The presence of the Participant 4 outlier (as mentioned earlier) may have had an undesirable impact on the test result, having skewed the overall representation of the "no experience" group. Additionally, it is important to note that each of the eight Mann-Whitney-Wilcoxon tests had only four data points per condition to base their results on. The drawback of a small dataset is a common issue in statistical testing and can lead to significant result distortions with the presence of otherwise uncommon outliers.

Conversely, the statistically significant test result for the time of Task 1 may highlight an important difference between participant groups: inexperienced participants initially needed to get used to the unknown interface. Seeing as Task 1 was the first that participants were asked to perform, time to get used to the UI for inexperienced users could have impacted the observed times. This explanation is also in line with observations that were made by the researchers: during the first task, participants with no experience "looked around" more than the experienced group. Nevertheless, though inexperienced participants took additional time to get used to the UI, they generally did not have any trouble finding buttons needed to accomplish the given task - this is also explained by the test results indicating the number of clicks performed being independent of their experience level.

3.4 Feedback Results

The feedback we received during the experiments can be divided into three sections: regarding features of the timed events window, the Statistical Overview page, and the application's buttons.

3.4.1 Feedback: Timed Events

Creating a clear separation between the usage of timed events (timer and stopwatch) is important to us. The feedback we received about how clear and intuitive this separation is indicated that most participants were content with the current toggle which visually separated the functionalities within a single window. Though some participants with no experience of the UI clicked a different button when asked to initialize a timer after having initialized a stopwatch, all of them agreed that the addition of a separate button for each would add unnecessary clutter to the UI; this separation is something that is easily learned after brief usage of the application as the toggle provided an intuitive enough separation.

Useful feedback regarding other items in the timed event window was also given. We should increase the size of the buttons in the window itself to make them more easily clickable, as participants struggled to do so with our hi-fi prototype. This is an easy fix for further iterations of the existing prototype. Additionally, it was noted by participants that the label for the timer slider was confusing. The hi-fi prototype gave the timer duration in minutes, which is a view that users are not commonly used to. Instead, this duration should be displayed in hours, as this is a more standard and recognizable format. For example, instead of displaying a two-hour timed as "120 m" it should be changed to "2:00h".

3.4.2 Feedback: Statistical Overview

From the feedback we gathered regarding the Statistical Overview page, we learned that the types of visualizations used (pie chart and bar chart) were easily approachable and readable by participants. Nevertheless, participants gave valuable feedback regarding stylistic choices of the interface.

The pie chart was easy to manipulate with the dropdown menu by all participants, but participants noted that the dropdown size should be increased, though its current placement is appropriate above the pie chart. Furthermore, many participants expected that the filtering of labels provided at the bottom of the page would also filter the labels shown in the pie chart; this is a logical assumption that will be provided in the future. The current placement of the label filters at the bottom of the page was also criticized, and participants suggested moving it further up on the page to make it more central and put it more in focus for users to find more easily.

We also collected valuable feedback regarding the appearance of the pie chart. Currently, the pie chart lacks explicit labeling of its constituent parts and requires referring to the actual label filters to understand what color refers to which activity. This can be easily fixed by adding explicit labels on the pie chart. Additionally, a participant pointed out that the blue color on the

pie chart made it difficult to read the text indicating the number of hours spent on that activity. Thus, we should either adjust the colors to provide better contrast with the font color or add a highlighting background-glow effect to the text to prevent such readability issues.

3.4.3 Feedback: Buttons

It was important for us to gather feedback about the buttons of the interface and its icons to understand how intuitive these represent their intended functionality. The button that seemed to be the least intuitive was the button to initialize a timed event, represented by a play icon. Many participants indicated that it seems more intuitive to use a clock icon instead, as this transmits a clearer idea that the button's functionality is related to time.

Aside from feedback regarding the button's icons, it was also suggested that adding textual labels to the buttons would enhance the understanding of each button's functionality. This is an interesting suggestion but requires further testing before a definite choice is made since this textual labeling may add undesirable clutter to the interface. A potential consideration is to make this an optional feature for users to enable themselves.

4 Conclusion

Our study aimed to measure how different levels of experience with our hi-fi prototype would impact participant performance, measured in the time and number of clicks required to accomplish a given task. From the experimental results discussed in Section 3, we can conclude that experience level has an impact on task completion time but may impact the number of interactions (clicks) required to navigate the UI.

First, we conducted a Mann-Whitney-Wilcoxon test for the total performance difference between conditions on both dependent variables. In the case of the Time(s) we rejected the null hypothesis and for the Number of clicks, we failed to reject the null hypothesis. Afterward, we conducted an additional eight Mann-Whitney-Wilcoxon tests for every combination of the dependent variable and task present, failing to reject the null hypothesis that there is no difference in the distributions of both groups for all tests except one for the time taken for Task 1.

From these conclusions as well as experimental observations made, we learned that our UI is intuitive and simple to use, even for inexperienced users. However, the statistical tests revealed that experience level does impact task completion time when users are initially presented with the new UI. Nevertheless, navigating the UI to perform tasks (measured by the number of clicks) seemed consistent across both experience-level groups. The integration of the timed event and statistics functionalities was well received and seemingly convenient to use by participants - a trend we hope will carry over to all users of our application.

4.1 Limitations

We identified several limitations with our study experiment.

Sample size Our study experiment consisted of eight total participants, four participants per test condition. This did not provide many data points for our study results and challenges the overall validity of the user study. As mentioned in Andrade's paper on the importance of sample size in research, "too small a sample is unscientific and also unethical" [1]. This highlights a major drawback in the significance of our results and the necessity for more extensive testing.

User population The population of our study consisted entirely of Computer Science students from the VU between the ages of 18-25. This is a clear limitation of our user study as this is only a subset of our application's user base. The narrowness of the type of participants we used for the study may have led to skewed results, likely guiding us to develop an application for a younger and more technologically adept subgroup. Ideally, future studies should sample participants more randomly for a more accurate representation of the target population.

Device familiarity The user study presented an unexpected issue: the user's familiarity with the laptop device they were provided, the Apple MacBook. The performance of Participant 4 was most notably affected by the usage of the device's trackpad. Their unfamiliarity with the MacBook's Force Touch trackpad caused significant issues with their completion of the provided task as the trackpad does not provide the same kind of haptic feedback as they are used to. This issue poses a limitation to the results that were gathered, as the impact of device usage had a more significant effect than anticipated.

Missing functionality In the hi-fi prototype we decided to only implement the functionality needed to support our user study. We ended up leaving a lot of important features: we left icons and explorable interfaces unimplemented. Therefore we cannot assume that the results found in our study are in correspondence with the results that would have been found had we implemented the full application. It could have been the case that inexperienced users might have taken more clicks in order to traverse the bookmark, but since this icon was unresponsive it only took an inexperienced user one click to find out that it was irrelevant for the task. Further note that we did not implement any pop ups when the unresponsive functionality was attempted, which in some cases resulted in additional confusion since the participants believed a misclick to occur, thus resulting in additional clicks and time taken. In any future studies we would like to implement the full application in order to accurately gather information and results about the usability of our application.

4.2 Future Work

The gathering of participant feedback has greatly benefited our outlook on potential extensions to our user study. To gain further insight into our application's usability and provide a more seamless user experience, we identified several interesting topics for future exploration.

Our application relies heavily on the use of icons to convey meaning about the functionality of each button. Thus, it is advisable to study what impact different icon usage has on the predictability of each button's functionality. Is the icon of a stopwatch preferable to that of a play button? These are the kinds of questions we should strive to answer.

A future study could not only test the type of icons used but also their placement and size in relation to other elements on the screen. For example, decreasing the size of the "Presets" button may put more emphasis on the other, more focal timed event button of the application.

Another major consideration in future studies is to accommodate a larger user audience. Seeing as not all people are knowledgeable about standard icons and their behavior, it should be tested whether adding textual information to describe each button's functionality enhances user understanding. However, how can we add minimal textual information while not cluttering the interface? It could be interesting to consider the addition of textual features as a customization of the application's UI that gives users the power of choice.

The ideas of future work presented are endless, but the focus on buttons and the icons used to represent them has been a trend in feedback that was gathered during this user study. We hope that future work can build on the ideas presented here to improve the usability of the application.

References

[1] C. Andrade. "Sample Size and its Importance in Research". In: *Indian Journal of Psychological Medicine* 42.1 (2020), pp. 102–103. DOI: 10.4103/IJPSYM.IJPSYM_504_19. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6970301/.

[2] Apple Inc. *How to use the Force Touch trackpad*. https://support.apple.com/en-us/102309. Accessed: 2024-03-23. 2023.

[3] Ben Shneiderman. *The Eight Golden Rules of Interface Design*. https://www.cs.umd.edu/users/ben/goldenrules.html. Accessed: 2024-03-23. 2016.

[4] University of Twente, Electrical Engineering, Mathematics and Computer Science (EEMCS). *Informed Consent Form Template for Research with Human Participants*. https://www.utwente.nl/en/eemcs/research/downloads-and-forms/informed-consent-form-template-for-research-with-human-participants-ec-cis.docx. Accessed: 2024-03-13. 2023.

Justification

Team Member	Responsibility
George	**Q1**: wrote design rationales **Q2**: wrote conclusion
Yasin	**Q1**: helped with file organization **Q2**: wrote method, helped with results & discussion, helped with introduction
Laura	**Q1**: organized all files, wrote README **Q2**: wrote results & discussion, helped with conclusion, helped with introduction
Tomáš	**Q2**: wrote abstract, keywords, introduction