User Study: Mental Tree (HCI 2024)

Cristina Ettlin
BSc Computer Science, ETH Zürich

Thiago Knill
BSc Computer Science, ETH Zürich

Khaled Kottmann
BSc Computer Science, ETH Zürich

Marcello Krahforst
MSc, Robotics, Sytems and Control, ETH Zürich

Noah von Matt
Bsc Computer Science, ETH Zürich

Marco Weder
Bsc Computer Science, ETH Zürich

Figure 1: Mental Tree, the app to uncover your mental load.

ABSTRACT

In our user study on the application "**Mental Tree**", we aimed to visualize the mental load of the end users. To achieve this, we developed an application that displays the mental state and current workload of each user within a household. In the study, we tested how well users could identify household members who might need help and how they could redistribute tasks by trading to reduce the mental load of that member. We gathered data using time measurements, number of miss-clicks, and System Usability Scale (SUS) questions, with our 12 participants chosen through convenience sampling. From our test results, we see that the aspects we initially aimed to measure did not significantly impact the collected data. Instead, we had to rely on the SUS scores, user feedback, as well as the observed user behavior to determine preferences. Overall, most participant preferred Prototype B, even though it was perceived as more confusing. Those who chose Prototype A generally did so because Prototype B was more difficult to understand at first glance.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Treemaps; Human-centered computing—Visualization—Visualization design and evaluation methods—Mental Load

1 INTRODUCTION

For our Course *Human Computer Interaction* in the Autumn Semester 2024 at ETH Zurich, we were tasked with creating an application to help people manage mental load.

Mental Load is the ongoing effort of organizing tasks and maintaining an overview to ensure everything functions smoothly, whether for projects or household responsibilities. Unlike the tasks themselves, mental load is often an invisible burden carried by a few people. This hidden effort cannot easily be delegated and can lead to stress or even depression.

A typical example involves a housewife or househusband who, regardless of how equally the physical tasks are divided, must always keep an eye on which tasks need to be done, when, and how tasks are related to each other. This means that the houseperson often has responsibility over all tasks, while other household members (such as the partner or children) may only notice if the tasks are being completed and by whom.

Our project focused on visualizing the mental load of users, resulting in the development of our application called **Mental Tree**. This is an application which is shared between all users of a household. At its core are the Mental Tree and a collection of flowers, with each flower representing a household member and easily identified by its unique color.

The flowers are visible to every user and can be displayed in three different states: Happy, Content, and Sad. Each user's *mood* is updated by the user themselves, providing an intuitive way to communicate their need for help or their current mental state to others.

The Mental Tree consists of multiple branches, each representing a category of tasks. Every branch also has a blossom for every user in different sizes. The Sizes represent the current workload for a user in each category. By pressing on a blossom, we can see more information about this category represented as a diagram and a list of tasks which the user can then trade, offer help, or remind others.

To make our Mental Tree effective, we needed a way to both track tasks and also distribute them among users. For this, we have implemented a playful representation of tasks as *Cards*, each containing information such as a name, category, image, priority (1 to 5), difficulty (1 to 5), due date, list of subtasks (names only) and notes. To interact with those cards, we have a *Cards Screen* which allows the user to see more information, create, edit, and delete cards, select their favorites by swiping each card to the left or right

and then submitting their selection. When every user has selected their task preferences, the cards get *shuffled* and distributed to each user according to their preferences. User are then able to check off completed tasks, undo them if needed, and view currently assigned tasks of other users to offer help, trade, or remind them.

For an overview of task progress and user moods, users can navigate to the *Diagram Screen*. Here, they can see each user's mood history, a pie chart showing how many tasks have been done by which user, a diagram representing the tasks done by each user in the last 20 days, the tasks done by each user sorted by categories and a List of tasks done and their check-off dates.

2 Study Design

The user study was designed to evaluate the high-fidelity prototype application, Mental Tree, which was developed using Flutter and runs on Android and iOS devices. For the testing we used laptops with Android device emulators. The application was modified to allow users to switch between Prototypes A and B, as well as to select between three different pre-programmed task sets from the settings menu. The prototypes differ in the diagram shown to the user when pressing a blossom on the tree. Prototype B shows a combined bar chart with both the number of completed tasks in the last 30 days and the number of all currently assigned tasks in the current shuffling period. Prototype A on the other hand, omits the latter and only shows the number of completed tasks in the last 30 days. The pre-programmed task datasets vary in size, with the small dataset containing an average of 5 tasks per category, the medium dataset contains an average of 10 tasks per category, and the large dataset averaging 20 tasks per user per category.

The prototypes are identically designed except for the diagrams shown when pressing on a blossom.

Independent Variables

- The diagrams shown in the two prototypes when pressing on a blossom in the Mental Tree

Dependent variables

- Time to complete a task (measured in seconds)
- Answers to a questionnaire based on the System Usability Scale test (SUS) containing questions on experience whilst using diagrams to complete tasks (only shown after doing all tasks with a given prototype)
- One question on task difficulty (Single Ease Question) (shown after each completed task)
- Manually encoded feedback from the users in inductive coding fashion

2.1 Hypothesis 1

- **H1**.**1**: For a small number of distributed tasks, the diagram presented in prototype A allows for faster task execution.
- **H1**.**1**$_0$: For a small number of distributed tasks, there is no difference in task execution speed between the two diagrams presented in prototypes A and B, or prototype B is preferred.
- **H1**.**2**: For a medium number of distributed tasks, the diagram presented in prototype B allows for faster task execution.
- **H1**.**2**$_0$: For a medium number of distributed tasks, there is no difference in task execution speed between the two diagrams presented in prototypes A and B, or prototype A is preferred.
- **H1**.**3**: For a large number of distributed tasks, the diagram presented in prototype B allows for faster task execution.
- **H1**.**3**$_0$: For a large number of distributed tasks, there is no difference in task execution speed between the two diagrams presented in prototypes A and B, or prototype A is preferred.

2.2 Hypothesis 2

- **H2**: The user will perceive the tasks completed with prototype B as more difficult than with prototype A
- **H2**$_0$: The user will find solving the tasks with both prototypes to be of the same difficulty, or find A to be more difficult than B.

2.3 Experimental procedure

We have used convenience sampling for our 12 participants. The users ranged in age from 20 to over 60, covering all genders, and from low to frequent smartphone users.

1. Before starting the study, the user is given some time to get familiar with the app and finish a small tutorial that explains all the relevant parts of the application the user needs to interact with.
2. The test subject is then assigned the task to offer a trade for a task that is assigned to him/herself with another user in the provided household in order to improve equity in the amount of tasks distributed among all users across all respective categories.
3. For each test subject, we alternate the order of tests and prototypes, to avoid introduction bias for the first task run. For each test we measured the time to complete the task
4. Every test subject runs through six tests consisting of every permutation between prototypes A and B along with the small, medium, or large datasets. This helps to get enough data for a small sample size of test subjects
5. After each completed task, the user reports the task difficulty through a single ease question
6. After completing every task, answer the SUS questions about the prototypes
7. Lastly, they can give a final feedback on the prototype preference after completing all tests

3 Results

3.1 Task Completion Time Measurements

3.1.1 Small Dataset

To compare the effect of the diagrams of the two prototypes on task completion time - specifically the task of trading a task with another household member in order to improve parity in task distribution among all users - in all respective categories, we conducted a Wilcoxon Signed-Rank test. Prototype A's data did not meet the assumptions of normality, whereas Prototype B's data did (Prototye A: $W = 0.713$, $p = 0.001$; Prototype B: $W = 0.866$, $p = 0.058$).

For Prototype A, the mean solving time was $M = 71.50$, $SD = 84.74$. For Prototype B, the mean solving time was $M = 68.00$, $SD = 49.51$.

The Wilcoxon Signed-Rank Test showed no statistically significant difference between the two prototypes in solving time, $W = 35.000$, $p = 0.791$.

These results indicate that there was no significant difference in solving time of the given task between the prototypes A and B for a small amount of tasks in the tree. Thus, we reject our Hypothesis **H1**.**1**.

3.1.2 Medium Dataset

To compare the effect of the diagrams of the two prototypes on the number solving time of the given task which is trading a task with another household member in order to improve parity in task distribution among all users in all respective categories, we conducted a Wilcoxon Signed-Rank Test, as the data from the prototypes did not meet the assumptions of normality (Prototype A: W = 0.729, p = 0.002; Prototype B: W = 0.743, p = 0.002).

For the Prototype A, the mean solving time was M = 51.83, SD = 51.06. For the Prototype B, the mean solving time was M = 70.25.00, SD = 69.99.

The Wilcoxon Signed-Rank test showed no statistically significant difference between the two prototypes in solving time, W = 23.00, p = 0.233.

These results indicate that there was no significant difference in solving time of the given task between the prototypes A and B for a small amount of tasks in the tree. Thus, we reject our Hypothesis **H1.2**.

3.1.3 Large Dataset

To compare the effect of the diagrams of the two prototypes on the number solving time of the given task which is trading a task with another household member in order to improve parity in task distribution among all users in all respective categories, we conducted a Wilcoxon Signed-Rank Test, as the data from Prototype A did not meet the assumptions of normality, in contrary to the data from Prototype B (Prototye A: W = 0.848, p = 0.034; Prototype B: W = 0.888, p = 0.111).

For Prototype A, the mean solving time was M = 44.42, SD = 30.84. For Prototype B, the mean solving time was M = 40.67, SD = 25.26.

The Wilcoxon Signed-Rank Test showed no statistically significant difference between the two prototypes in solving time, W=35.000, p=0.824.

These results indicate that there was no significant difference in solving time of the given task between the prototypes A and B for a small amount of tasks in the tree. Thus, we reject our hypothesis **H1.3**.

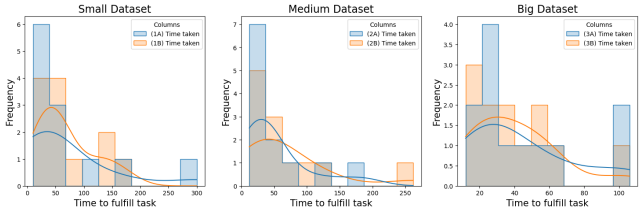

Figure 2: Distribution of time to fulfill task by amount of tasks distributed in Mental Tree

3.2 System Usability Score (SUS)

- The SUS score form prototype A is 70.57291666666666
- The SUS score form prototype B is 61.197916666666664

These results indicate that the users preferred prototype A over B.

3.3 Single Ease Question

To compare the effect of the two different diagram representations on the perceived task difficulty reported by the users after completing the given task, i.e. balancing the task distribution within the categories, we first check if the data is normally distributed.

Running the Shapiro-Wilk Test on the reported difficulties of the tasks for each prototype reports that except for the perceived difficulties of the tasks done with the prototype B with the small and big dataset, none of the distributions is likely to be normally distributed (Prototype 1A: W = 0.843, p = 0.030; Prototype 1B: W = 0.894, p=0.133; Prototype 2A: W = 0.757, p = 0.003; Prototype 2B: W = 0.779, p = 0.005; Prototype 3A: W = 0.809, p = 0.012; Prototype 3B: W = 0.877, p = 0.080)

Thus we run the Wilcoxon Signed-Rank Test in order to evaluate if there is a statistically significant difference between the two perceived task difficulties when solving them with each prototype. The test results in no statistically significant difference between the any of the groups. (1A vs 1B: W = 12.00, p = 0.201; 2A vs 2B: W = 8.00, p = 0.589; 3A vs 3B: W = 14.00,p = 0.566). The mean difficulties for each task can be retrieved from figure 3.

Overall it was perceived to be fairly easy to solve the task with both the diagram from prototype A and the diagram from prototype B. The mean difficulty score for prototype A was 3.94 across all tasks (with 5 being very easy and 1 being very difficult), with standard deviation of 0.127. Solving the tasks with prototype B was perceived as slightly more difficult with a score of 3.69 and a standard deviation of 0.256.

Due to the large variance in comparison to the measured difference in the mean value and thus the failed Wilcoxon Signed-Rank Test, the Hypothesis 2 can be rejected.

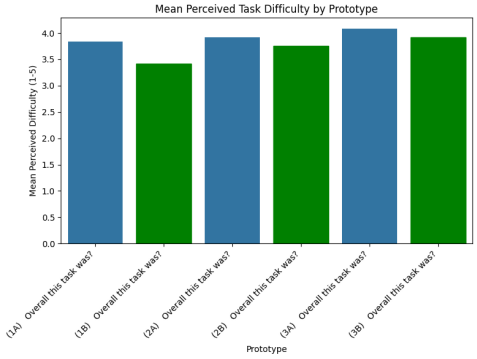

Figure 3: Comparison of the mean perceived task difficulty; (1A) stands for small dataset task with prototype A

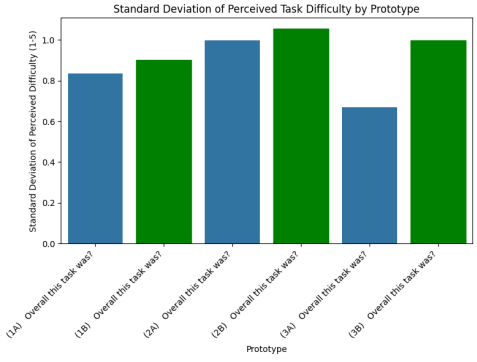

Figure 4: Comparison of the standard deviation in perceived task difficulty; (1A) stands for small dataset task with prototype A

3.4 Results from the qualitative questions/feedback:

The Encodings made were made by two independent testers and came to the conclusions in the following.

Prototype B was mostly preferred because it conveys more information and allows for better planning, also showing information

about the future. Users suggested splitting the graph into two graphs and potentially changing the graph type to convey information more clearly

Prototype A was mostly preferred because of its simplicity. By omitting confusing elements, it seems more welcoming to the user and less overwhelming.

It is interesting to note that more people preferred Prototype B (7 out of 12) but the SUS score states that Prototype A was preferred

4 Discussion

From our results, we can see that our hypotheses **H1.1**, **H1.2**, **H1.3** and **H2** were rejected.

For our Hypothesis 1 this means, that the additional information in Prototype B of showing the currently assigned tasks in the diagram as well, had no effect on completing the task of figuring out which tasks were good to trade.

Nonetheless, one can observe that the users are actually preferred prototype B over A according to the feedback, however the SUS hints that A is better than B. The reasoning behind this can be achieved from the qualitative questions and feedback. Although many were confused by Prototype B, they preferred it over A, since it shows relevant information. This would suggest that those participants chose B, but in the SUS had a tendency for A. Thus it might be worth experimenting with providing the user with the information given in Prototype B but modifying the way the information is presented to the user.

Finally, Hypothesis 2 can be rejected since there was not enough of a difference in the data to decern the resulting difference from noise. Even if one would disregard the deviation, the mean perceived difficulty shows a constant edge towards the Prototype A which further underlines that the prototype B did not perform as well as predicted, since it seems to have made the task more difficult for the users, adding to the received feedback of feeling overwhelmed by the visualization.

It should also be noted that we could measure a clear decrease in both task completion time and perceived difficulty as the participants progressed through the tasks. This further brings the limitation of the within subject testing approach to light which brings the inherent issue that the users learn to get better at the given task over time. However, this also shows that the prototypes are intuitive enough such that the user is able to learn to use the functions in a rather short amount of time.

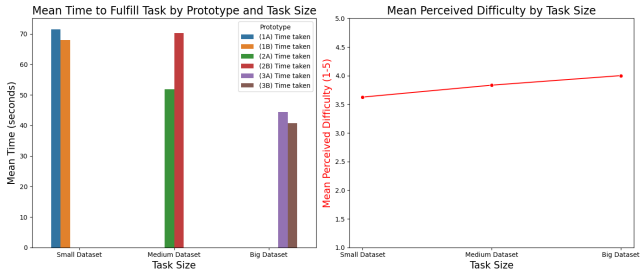

Figure 5: Decreasing time to solve and perceived difficulty (more means easier) as users solve task with different datasets

5 Limitations

In our test we used convenience sampling to choose our participants, which is not representative for all demographics, and thus our results might be partially biased. Since our sample size only consists of twelve participants, our data does not seem to follow any predicted structure and is instead heavily influenced by the noise of outliers, making it hard to see how the data would converge with a larger sample size.

Since our tests were very repetitive in a short period of time, our users experienced recency bias and got faster with each test. However, we tried to cancel the increased time measurements out by varying the prototype with which we start testing among the users. Because of this, our collected data on the taken time would be more representative as a learning curve of our User-Interface instead of the usability for the A-B prototypes with each size of datasets.

Finally, it should be stated that each of us ran the tests on a different device which might have an influence on the user experience which we did not control for.

6 Future Work

There were many users which were confused by aspects of our design. The most prominent was the additional feature in Prototype B. Showing two different types of information within one graph not only confused users, but also discouraged them to find out by themselves what their meaning was. With a more simplified design, Prototype B might be the preferred choice, because most of the participants preferring Prototype A did so only because Prototype B was more confusing to understand.

Furthermore, the combination of limited sample size, within-subject testing and convenience sampling probably skewed our collected data. It would certainly be beneficial to repeat the study with more participants, a between-subject testing approach and sampling a broader demographic. With more participants one could also vary the order in which the task datasets are tested, further removing the skew introduced by the participants getting better at doing the task.

Another important aspect, however, was seen from the behavior of the participants. When trading tasks and changing between the prototypes, the users did not acknowledge the diagrams. On multiple occasions, when filling out the SUS questions, users were confused about the differences between both versions.

Additional feedback and user behavior hinted at other aspects from our application, which could be enhanced. Participants often tried to click on Category names instead of the blossoms, which caused confusion when nothing happened. In our Tests, Users were encouraged to help out users with many tasks, causing them to click on the largest blossoms. Some users however, tried to choose a smaller blossom and had to press multiple times until it registered.

7 Conclusion

The conducted test on task completion time has resulted in the rejection of our hypotheses that prototype B would improve the usability of the app and thus the efficiency of the users. As for the preferences of the diagrams shown in the two prototypes, the SUS score indicates that Prototype A was better received, however the majority of the users, seven out of twelve, stated that they prefer Prototype B over Prototype A. The conflicting results from the user feedback and the SUS score indicate that there is potential in the idea of providing the user with more information, but it should be implemented differently.

Acknowledgments

The authors wish to thank Dr. Yi-Chi Liao, Zhipeng Li, and Rachel Schuchert for their help.