Reinventing Calendar Apps using Templates

Adam Dada, Rafael Fernandes, Rahul Kaundal, Oliver Lehmann, Matthias Lott, Moritz Mani

ETH Zürich

1 Introduction

As part of the HCI course at ETH Zürich, we applied the human centered design pipeline. From needfinding and ideation to three iterations of prototypes, we identified a problem that current users have with calendar apps, brainstormed solutions and came up with two prototype variants that were compared in a user study.

The main requirement that users demand in a calendar is that interactions should be swift and effortless. Several iterations of prototypes revealed that this problem can be solved using templates.

Templates are a special kind of event state that already contain some information such as time, location or participants. They expedite the event creation process, requiring only the completion of remaining empty fields to create an event. For example, a doctor's appointment will most likely happen at the same location and can be specified by a template.

Interacting with the templates is very important. Every time a user wants to create an event or modify a template, they are prompted to select a template. Thus, it is essential to get the design of the template selection screen right. For this purpose, we came up with two variants (Figure 1):

- *Variant A*: This variant arranges items in a grid which provides higher information density. Templates are displayed in two vertical rows of small squares that contain metadata such as date, location, timing and more.
- *Variant B*: This variant shows a vertical list of full-width template items holding their respective metadata. It makes the different templates more digestible and the screen less crowded with information.

Apart from having a different template arrangement, both variants are identical. This was ensured by implementing them both in the same application and adding a toggle in the settings menu to switch between them at will.

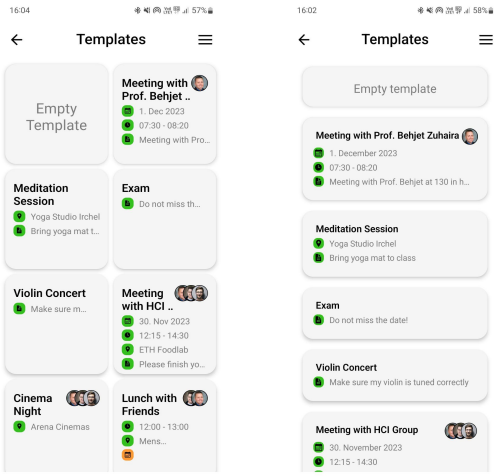

Figure 1: Template Selection in variant A (left) and B (right)

2 Study Design

The study was designed to analyze how the layout of the template selection screen influences the nature, ease and efficiency of the users' interaction with the prototype. We came to this decision because we had imagined the screen layout to play a pivotal role in the event creation process as the number of stored templates increased. Furthermore, since the screen for selecting templates is used for template editing as well as creating, finding the correct layout would yield significant reward in terms of speed and usability when performing certain user actions.

2.1 Study Setting

There was no unified hardware used as all studies were conducted on the personal devices of the group members (running either iOS or Android). Similarly, there was no consistent environment for the participants. Many studies were conducted at home, some of them were done remotely. However, we made sure to strictly abide by the study protocol [1].

2.2 Variables and Hypotheses

The only independent variable is categorical and encodes the interface variant (A or B). To ensure that memorization ability is not an independent variable, the list of templates was shuffled every time the user accessed it.

The study analyzes four dependent variables:

1. *Actual Speed*: The time that a user spent on the template selection screen. It is an objective interval variable.
2. *Perceived Speed*: The time a user thinks they spent on the template selection screen. It is the subjective counterpart to actual speed and also an interval variable.

3. *Scroll Amount*: The amount a user scrolled while looking for the correct templates. It is an objective interval variable.
4. *Usability*: This indicates how usable the template system is and was captured using the system usability score (SUS) questionnaire [2]. The answers were aggregated into a total usability score between 0 and 100. The answer to each question is a subjective ordinal variable, the final score is too.

These four dependent variables naturally induce the following null hypotheses:

1. There will be no effect of changing the grid view to a list view on the time a user needs to select a template (actual speed).
2. There will be no effect of changing the grid view to a list view on the time a user thinks they need to select a template (perceived speed).
3. There will be no effect of changing the grid view to a list view on the amount that the user has to scroll (scroll amount).
4. There will be no effect of changing the grid view to a list view on the usability of the template system (usability).

2.3 Procedure

Participants were selected using a method called convenience sampling [3]. Information about the participants can be found in section 2.4.

To test whether the null hypotheses hold, a task consisting of two subtasks was devised:

1. Edit the "Lunch with Friends" template by setting its color to green.
2. Plan a trip to the zoo using the template "Zoo" next saturday lasting the whole day.

To compare the two design variants and find out if the hypothesis can be rejected, a counterbalanced within-subject study was conducted. Every participant tests both variants, half of them starting on variant A, the other half on variant B. The study consists of the following steps (further details can be found in the Study Protocol [1]):

First, the user fills in a questionnaire in which *participant data* (age, gender, technical affinity and familiarity with digital calendars) is anonymously collected. Subsequently, they are presented with the task and asked to practice it once on both variants. After the trial run, they perform the task again on the first variant during which we collect *performance data* (time spent selecting a template and vertical scroll amount). Afterwards, the user fills in a questionnaire that collects *subjective ratings* (perceived time spent selecting a template and usability). This process is then repeated for the second variant. During the whole study, *qualitative data* (verbal feedback from the user and other observations) is gathered and discussed with the participant at the end.

2.4 Participant Distribution

The twelve participants that were interviewed for this study are peculiarly distributed. Most participants were either fairly young or fairly old. Collectively, they had an average age of 44.333 years (median: 50.5 ys, SD = 21.073 ys). Their gender was more balanced: 7/12 participants identify themselves as female, 5/12 as male and 0/12 as other/none. Their technical affinity was distributed similarly to their age: Six people rated themselves a 4 out of 10, three people each rated themselves at 8 and 9 respectively (median: 6, SD = 2.379). The frequency with which they use calendars was also quite strange: Only one person uses a calendar app daily, 4 people use a calendar app often, the remaining 5 people use a calendar app only rarely or never.

3 Results & Discussion

3.1 Actual Speed

To compare the effect of the **interface variant** on the **time spent on the template selection screen**, we conducted a **paired samples t-test** (all Shapiro-Wilk p>0.05 and Levene's p>0.05).

With variant A, participants on average needed 40.837 seconds (SD = 12.935 sec) to select the templates. With variant B, participants on average needed 46.132 seconds (SD = 16.365 sec) to select the templates (see Figure 2).

The mean difference between the two groups was **not** statistically significant; t (11) = -1.384, p=0.194.

There is no indication that template selection on variant A is faster than on variant B. H0 is **not** rejected. A reasoning for this might be that the two variants have two different factors that hinder user speed by a similar amount. Firstly, variant A had participants scan the templates in a zig-zag fashion instead of a straight line. Secondly, variant B has templates displayed less densely, requiring more eye movement overall. This is supported by a user remarking that they were slowed down by the zig-zag reading.

3.2 Perceived Speed

To compare the effect of the **interface variant** on the **time participants think they spent on the template selection screen**, we conducted a **paired samples t-test** (all Shapiro-Wilk p>0.05 and Levene's p>0.05).

With variant A, participants thought they needed 6.167 seconds (SD = 3.486 sec) on average to select the templates. With variant B, participants thought that theyneeded 7.167 seconds (SD = 4.174 sec) on average to select the templates (see Figure 3).

The mean difference between the two groups was **not** statistically significant; t (11) = -1.864, p=0.089.

There is no indication that users perceive template selection as faster on interface variant A compared to variant B. H0 is **not** rejected. The two variants may be perceived similarly, because users are performing the same actions in a similar intensity (scrolling, reading and tapping). However, because we observed a p-Value of 0.089 which is quite close to 0.05, we also provide a different interpretation: In case a user never has to scroll, there are

indeed less actions to complete (no scrolling). As variant A has a more dense visualization (four times no scrolling was needed, in variant B this only happened twice) the perceived speed might be slightly lower.

As a note, *perceived speed* is lower than *actual speed* because firstly, testers probably did not consider app response-time as part of the perceived speed and secondly, testers likely did not include the time it took to navigate from the detailed template summary to event creation/template modification which was recorded in the actual speed metric.

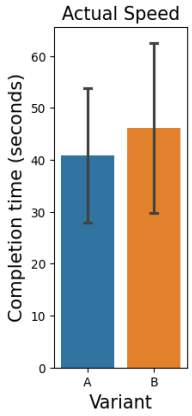

Figure 2: Actual Speed

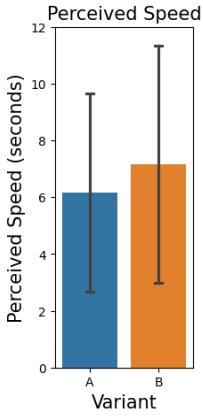

Figure 3: Perceived Speed

3.3 Scroll Amount

To compare the effect of the **interface variant** on the **amount that the user scrolled on the template selection screen**, we conducted a **wilcoxon signed-rank test** (Shapiro-Wilk (Variant A) p>0.05, Shapiro-Wilk (Variant B) p<=0.05 and Levene's p>0.05).

With variant A, participants on average scrolled 1448.417 pixels (SD = 1064.636 px) while selecting the templates. With variant B, participants on average scrolled 2116.583 pixels (SD = 2117.855 px) while selecting the templates (see Figure 4).

The mean difference between the two groups was **not** statistically significant; Z = -0.863, p=0.388.

There is no indication that users scroll less on variant A than on variant B. H0 is **not** rejected. A reason for this might be that users feel less secure on variant A and start asking themselves if they missed the template and scroll back to already visited sections to recheck whereas in variant B this is not the case. A participant remarked that variant B is calmer and the reading flow goes cleanly from top to bottom.

3.4 Usability

To compare the effect of the **interface variant** on the **usability (Total SUS score)**, we conducted a **wilcoxon signed-rank test** (ordinal data).

With variant A, participants on average rated the usability of the template system with 85.833 points (SD = 13.831). With variant B, participants on average rated 83.333 points (SD = 12.123) (see Figure 5).

The mean difference between the two groups was **not** statistically significant; Z = -1.358, p=0.174.

There is no indication that variant A is more usable than variant B. H0 is **not** rejected. Users rated the usability of the interface variants almost identically, this might be because they simply don't differ much. After all, users still perform the same actions (scrolling, reading and tapping). It is worth stating that a user mentioned they like that they can see more information at a glance in variant A.

According to [2], a score of 68 means that usability is average. With a usability above 80 in both interface variants, the template system is indeed quite usable.

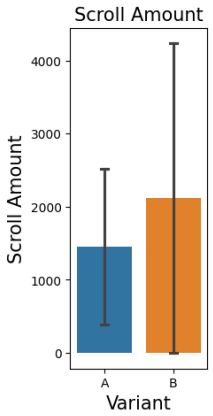

Figure 4: Scroll Amount

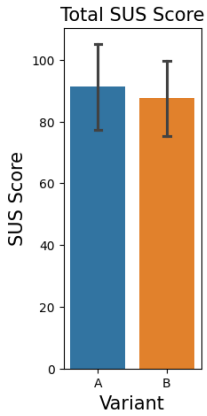

Figure 5: Total SUS Score

3.5 Qualitative Feedback

The qualitative feedback concerning the dependent variables we measured was already discussed in the respective sections above. Additional feedback given by the participants is presented in the following paragraphs:

Some users were confused whether their template changes were actually applied. An explanation for the confusion could be the lack of visual reassurance that changes were made permanent. In addition, study participants did not immediately understand that they can create an entry that is not based on a template using the first element in the template collection. This might be caused by the pressable "Empty Template" button being too visually similar to the rest of the items and, hence, did not stand out as having a distinct function.

The verbal feedback of the user study participants also highlighted the fact that the green checkmark in the top right corner of the screen when editing events or templates was falsely implying auto-saving functionality. Instead of representing an icon button which saves all changes that were made to the current entry or template, many users believed the green checkmark to be a confirmation that their changes were saved. A text button with an instructive label such as "Done" or "Save changes" would, according to testers, have avoided this ambiguity. Another feedback that relates to the screen for editing or creating entries was the overload of icons. Users reported difficulties in differentiating visual-only icons from functional ones.

3.6 Effect of Participant Distribution

As already mentioned in section 2.4, the distribution of participants was not as expected. Due to the size constraint of this report, we are not able to discuss the distribution of participants in much detail but we noticed a tendency that *technically affine* people prefer variant A.

If we run the analysis again, however this time merely for the six people that think of themselves as technically affine, we obtain p-Values 0.055, 0.028, 0.345 and 0.039, indicating a statistically significant difference in *perceived speed* and *usability*. There was also an almost statistically significant difference in *actual speed*.

A possible reason for this might be that technically affine people tend to be more comfortable using electronic devices. They experience less mental load handling the device and use this excess mental capacity to process the denser information in variant A, leading to faster task completion, less scrolling and higher usability.

Running the analysis again for the less technically affine people led to very high p-Values. This lets us believe that they either cannot take advantage of variant A and are indifferent to the interface variant, or might even prefer variant B.

4 Limitations

A profound limitation of the user study was that only a debug build of the prototype was used. There were problems exporting the app and, as a result, the performance was not ideal. For example, during template selection, the order of the templates was shuffled each time a user was supposed to select them. However, sometimes the screen was shuffled only after a delay, impacting the *actual speed*, *perceived speed* and *scroll amount*. This leads us to believe that the results of the study might not be applicable to the participants.

Another considerable restriction was the number of participants in the study. Having only twelve participants makes it very hard to analyze results (detecting outliers, concluding normally distributed data, etc.). Furthermore, the results of the study are not generalizable to a broader population in our opinion. This is, on the one hand, due to the small sample size and, on the other, due to participants (family members of young adult ETH students) not being representative of the wider population.

Furthermore, there was no unified study environment which led to the addition of several independent variables. Different participants were interviewed in different locations, on different and unfamiliar devices with varying screen dimensions that impacted scrolling amounts, and during different times of day. These factors might be a good explanation for the high standard deviation of the dependent variables.

Additionally, despite the fact that sufficiently many templates were used to make the screen scrollable, having even more templates might have led to a bigger difference between the variants.

5 Future Work

To begin with, we would like to redo the user study. This time, we would aim to have more participants and be more selective in choosing study participants, ensuring enough technically affine and non-affine users (ideally also some people that rate their affinity to be average). As explained in section 3.6, this should allow us to gain insights into the preferences of different user-groups and enable us to make a better-informed decision as to which interface to keep.

Furthermore, we would implement the improvement suggestions voiced by the participants of the user study. Such changes include the usage of a production build to improve the performance of the prototype, visual feedback (e.g. a snackbar) that confirms changes to events or templates**,** color coding templates and events according to their assigned color, and a floppy disk icon or a “Save” text button instead of a green checkmark to make the saving of changes on the edit screen very clear.

As it relates to the selection of dependent variables, we would additionally measure the mental load and scrolling *behavior* of study participants for both variants. This would shed more light into why users prefer a certain variant and how these variables correlate among different user-groups.

Lastly, we would enforce a unified environment to conduct the user study in. As a means to mitigate the various independent variables that were unintentionally introduced we would conduct the prototype testing for all participants on the same device, in the same location and at similar times during the day.

6 Conclusion

This report analyzed the impact of different template arrangements on a user’s ability to select templates quickly and easily. In a user study, four variables (actual template selection speed, perceived template selection speed, scroll amount and usability) were examined for two interface variants (grid view (A) and list view (B)). No statistically significant difference was found between them. This was accredited to a non-production app, a limited number of participants, different study environments and not having enough templates. However, we noticed that by restricting ourselves to only technically affine people, the results changed drastically. As a next step, we would like to implement the feedback we received from users as mentioned above, improve our study and conduct it again on a larger set of participants with a more diverse technical affinity.

References

[1] HCI Group 12. Study Protocol. https://github.com/eth-hci-course/hci-project-hci2023-group-12/blob/2023/Deliverables/Study/Study_Protocol.pdf.

[2] U.S. General Services Administration. System Usability Scale (SUS). https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html. Accessed December 2023.

[3] Wikipedia. Convenience sampling. https://en.wikipedia.org/wiki/Convenience_sampling. Accessed December 2023.