Study Report A/B testing Group 16 LunchSphere

Andri Bernhardsgrütter Charles Kremer Nishanth Kumar Alex Schlieper Alex Staikov
Nicolas Stucki

1 Introduction

In this study we aim to assess the impact of a swiping interface on accepting group meal appointments in our LunchSphere app. We have chosen to conduct an A/B test specifically on this feature because we believe efficiency and simplicity are key qualities for a planning app. Additionally, the selection feature is the task performed most often as it is required by every user every day. Our first approach for group selection is a vertical scrolling interface. We are also interested in a second interface that is based on swiping, similar to the well known Tinder dating app. We suspect that the swiping action improves user satisfaction, which is essential for user retention. However, the swiping feature might come at the expense of selection speed. Specifically, we are interested in understanding how the swiping feature influences the user's System Usability Scale (SUS) score, the time required to select a group, and the number of interactions (number of screen touches) users need to perform.

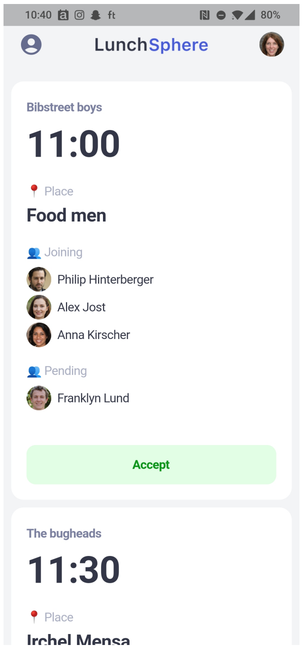

Figure 1: Interface A: List-Based Lunch Scheduling

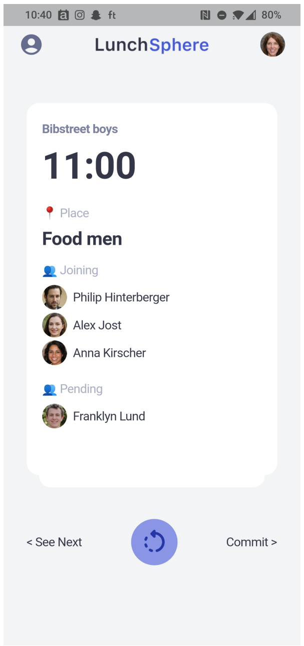

Figure 2: Interface B: Swiping-Based Lunch Scheduling

2 Study Design

Study Setting: The apps were installed on our smartphones (2 Android and 4 iOS devices) to present users with an environment similar to actual usage. Each study was conducted individually at various locations by our group members. The study protocol was followed, with data being collected both by the app programmatically (task execution time and number of interactions) and through post-task questions.

Independent Variable: The independent variable is the lunch scheduling process. To perform the study, we have developed two different prototypes (A and B). They display nearly identical group information cards with details such as time, place, and participating members listed vertically. Additionally, we include people listed as pending, which indicated that they are part of the group but have not accepted yet. The "accept" action commits the user to the selected lunchtime and it is implemented differently in A and B. Prototype A lists the cards vertically and adds and Accept button to the card. Protype B implements the Swiping feature where the cards are displayed as a stack with the top one visible to the user. The user can swipe left to be presented with the next card, swipe right to accept or use the unswipe button to see the previous card. The unswipe button is located centrally at the bottom along side text that reminds the user which direction corresponds to which action. Fig. 1 and Fig 2 show the respective prototypes.

Dependent Variables: Three dependent variables were measured in our test:

- **SUS Value:** A straightforward ten-question questionnaire assessing subjective system usability [1].
- **Interaction Count:** The number of times the participant touched the screen, specifically the number of time a finger was released from the screen. A button press and dragging/swiping motion all count as 1 motion respectively.
- **Task Completion Time:** The duration between opening the selection page to committing to a lunch group. Selecting the A or B prototype is performed by the user by pressing a button in a prototype selector page.

Hypotheses:

- **NH1**: The use of a swiping-based selection, in comparison to a list-based selection, has no impact on the SUS scores for user experience.
- **NH2**: The use of a swiping-based selection, in comparison to a list-based selection, has no impact on group selection time
- **NH3**: The use of a swiping-based selection, in comparison to a list-based selection, has no impact on interaction count to select a group.

Experimental Procedure: Study participants were gathered by convenience sampling. The study was organized as follows.

1. Introduction to the Experiment.
2. Training Phase: Show the participant a pre-recorded video with explanations of the corresponding prototype.
3. Provide the participant with an imaginary lunch scenario and let them select a group that meets the given criteria. The criteria were time constraints, preferred cafeteria, and people that need to be avoided. The scenario was read out but a cheat sheet with necessary constraints was provided for convenience.

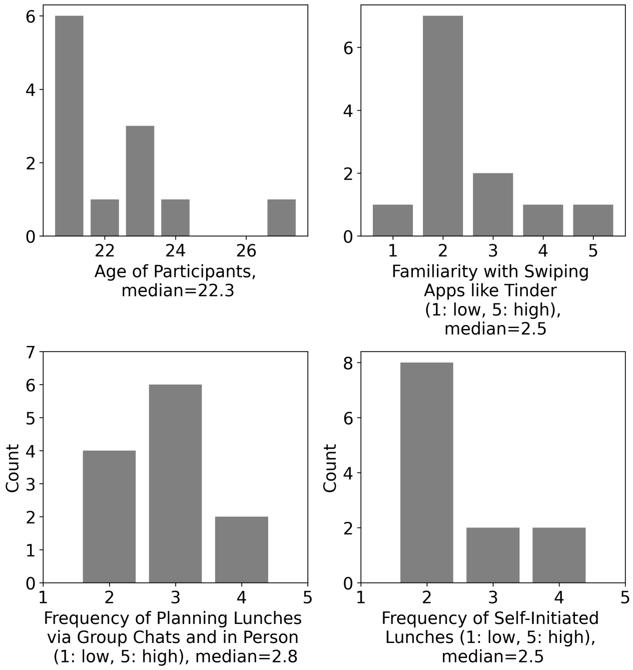

Figure 3: Visualization of pre-questionaire results.

4. Retrieve task completion and interaction count.
5. Ask the participant to complete the SUS rating scheme.
6. Conduct a post-task interview with two questions. (”What were your thoughts when using the app for scheduling?” and ”Were there any aspects of the app that confused you or that you found difficult to use?”)
7. Switch prototypes (return to the training phase and repeat the steps).
8. Finally, ask the participant to compare both versions and express their preferences.

Participants: Figure 10 visualizes the key data. The 12 participants had a mean age of 22.3 (SD=) and a gender balance of 50/50 exactly. The more task specific demographics were first: Familiarity with swiping apps like Tinder. There the majority of participants were not really familiar. Then: Frequency of planning lunches via Group chats and in person. This was very mixed and most people plan their lunches both in chats but also in person. Then we asked how often the lunch plans are self-initiated and most people rarely self-initiate it. The pre-study questionaire is summarized in Figure 3.

3 Results

The figures 4.-9. show histograms of the quantitative study results.

3.1 Quantitative Results

NH1: To compare the effect of the selection method on the SUS value, we conducted a Wilcoxon Signed Rank Test (all Shapiro-wilk $p < 0.05$ and Levene’s $p > 0.05$). Participants presented with prototype A gave an average SUS rating of 90.0 (SD =8.0) whereas prototype B received an average SUS rating of 82.9 (SD = 11.0). Both SUS means are high, indicating that both prototypes are effective in terms of usability. The mean difference between the two groups is statistically insignificant (Z = 13.0, p = 0.14). This means that there is not enough statistical evidence to reject NH1. From a usability standpoint as measured by the SUS, there is no discernible difference between the two interface designs. A possible interpretation is that both interfaces are equally effective in terms of usability, implying that users find them similarly user-friendly, efficient, and satisfactory. One could alternatively also conclude that he SUS is not sensitive enough to detect subtle differences between these particular interfaces or that the sample size was too small (N=12).

NH2: To compare the effect of the selection method on task completion time, we conducted a paired samples t-test (all Shapiro-wilk $p > 0.05$ and Levene’s $p > 0.05$). With interface A, participants, on average, needed 22.3 seconds (SD = 12.7 sec) to complete the task. With interface B, participants, on average, needed 30.4 seconds (SD = 21.0 sec) to complete the task. Again, the difference is not significant $t(11) = -1.8, p = 0.11$. The absence of a significant difference suggests that both interfaces are equally efficient in enabling users to complete task. It implies that choosing one design over the other will not detrimentally affect the efficiency of task completion. However, it might also indicate that the tasks tested were not sufficiently challenging or varied to elicit differences in performance between the interfaces.

NH3: To compare the effect of the selection method on gesture count, we conducted a paired samples t-test (all Shapiro-wilk $p > 0.05$ and Levene’s $p > 0.05$). Participants presented with prototype A needed 13.4 gestures on average (SD = 5.8 gestures) to select the right group. The same task required on average of 9.6 gestures (SD = 4.1 gestures) to complete the task. The mean difference between the two groups is statistically significant ($t(11) = 2.49, p = 0.03$). Given this result, we have enough statistical evidence to confidently reject the null hypothesis assuming a significance level of 5 percent. It suggests that the design of prototype B enables users to complete the task with fewer interactions. This could have important implications for the user experience, potentially indicating a more streamlined or user-friendly interface in prototype B. A swipe is a definitive complete action, a scroll action can be infinitesimally small. Therefore interaction count does not entirely relate to usability.

Interestingly, we observed an inverse relationship with task completion time. While prototype B necessitated fewer interactions on average, it’s crucial to consider the nature of these interactions. This distinction underscores that not all interactions are equivalent in their contribution to usability. Therefore, the lower interaction count in prototype B does not straightforwardly translate to higher usability. It suggests that while prototype B’s design facilitates task completion with fewer gestures, these gestures might be more significant in terms of user effort or decision-making compared to prototype A.

3.2 Qualitative Results

To receive qualitative results, each participant was asked the same two questions after each prototype. (”What were your thoughts when using the app for scheduling?” and ”Were there any aspects of the app that confused you or that you found difficult to use?”). We summarize the most important points below.

Prototype A: Approximately half the participants mentioned that they enjoyed the overview one has with a scroll based interface. About a third of the participants were unsure whether the lunch groups were ordered and, if so, according to which criteria. The lunches were in fact ordered according by time, but that had not specifically been explained. In a similar fashion, some mentioned that a filter feature would be useful.

The idea of A’s better overview mentioned by many users could be caused by several things. Firstly, scrolling allows the user to show parts of two different cards at the same time, thereby making comparison between groups easier. Secondly, many users were not as familiar with the swiping interface as we expected, suggesting that it took some cognitive load to learn the mechanic. It should however also be noted here that the question ”Familiarity with swiping apps

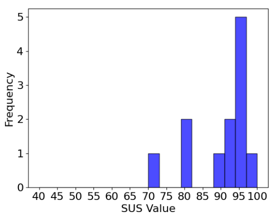

Figure 4: SUS values A version

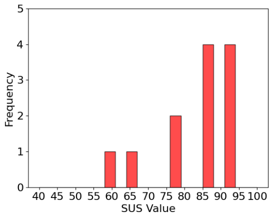

Figure 5: SUS values B version

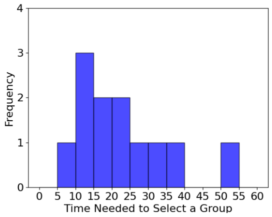

Figure 6: Task completion time A version

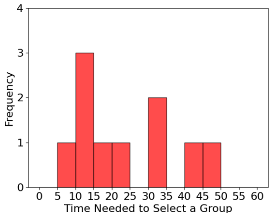

Figure 7: Task completion time B version

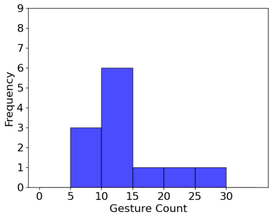

Figure 8: Gesture count histogram A version

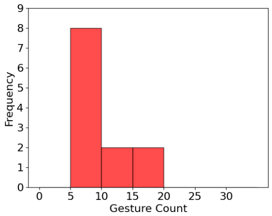

Figure 9: Gesture count histogram B version

like Tinder" is ill chosen, as it might be embarrassing for some to answer truthfully.

Prototype B: About a third of the participants mentioned that they found the swiping prototype fun to use. Approximately a fourth of the participants were partly confused by the swiping text indications at the bottom of the screen, as some confused them with buttons. A few participants stated that the swiping feels unintuitive or inappropriate for a lunch planning app. Similar to prototype B, some participants stated that they were unaware of the chronological ordering.

Comparison: At the end of the study, each participant was asked which prototype they preferred and why. Two thirds of the participants preferred prototype A, a sixth of them favored prototype B and the remaining could not decide. The provided argument was generally the overview of A against the more 'fun' to use version B.

4 Discussion

In this study, we explored the impact of two different interface designs on usability in the LunchSphere app. We suspected to find a tradeoff between task efficiency and satisfaction, but the results are not significant for either side. The SUS score of the two versions is very similar. The qualitative results indicate us that while most preferred A, many found B more enjoyable. Unfortunately this study could not show a significant difference in task completion time, which we believe would be the most important finding of the study. The study would have to be repeated with a bigger sample size. One notable observation is the significant reduction in interaction count required in the swiping interface. We were surprised by this inverse relationship to task completion time, but as discussed in Section 3.2, interaction count does not entirely relate to usability.

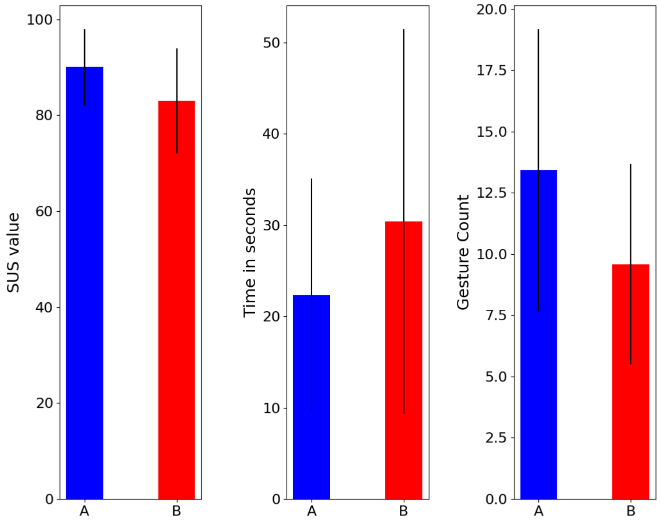

Figure 10: Mean and standard deviation from left to right for SUS value, task completion time and gesture count

The study unfortunately does not consider cognitive load. It would be interesting to expand on our suspicion that a swipe is mentally more draining than a scroll action. Since seeing the previous card is more time consuming in B than A, the user puts more effort into remembering information from previous card.

5 Limitations

The study's primary limitation was its small sample size, which likely lead to a failure to show significant results for NH1 and NH2. While the study's validity is limited to the age category of most samples, we do not believe this to be inaccurate as our primary target audience is university students. This is however different for their cultural background, as all participants are residents of Zurich. Additionally, the experimental setting, which is very similar to but still differs from real-world usage scenarios, might have influenced user behavior. Of course, an improved study would analyze behaviour of real world users once they use the app daily.

6 Future Work

Future work for this study could focus on several areas: exploring how a filter function might change user interaction and decision-making, investigating the influence of the order in which group information is displayed, evaluating the necessity of showing pending member information, and applying these findings to real-world scenarios.

7 Conclusion

The study does not show significant results to confidently decide between both versions. Through qualitative results, we are inclined to use prototype A for the first version of LunchSphere. We plan to repeat the experiment with a larger userbase.

References

[1] J. Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.