# Study Report A/B testing Group 16 LunchSphere

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

## 1 Introduction

The aim of this study is to assess the impact of a swiping interface for accepting group meal appointments in our LunchSphere app. Specifically, we are interested in understanding whether this swiping feature affects the User's System Usability Scale (SUS) score, the time required to select a group, and the number of interactions (taps/swipes) users need to perform. We have chosen to conduct an A/B test on this feature as we believe that simplicity and efficiency in planning are key qualities for a planning app, based on feedback received. In the A version of the study we have a listing of lunch groups you can join with the time, place and people that are joining listed. Also there is an accept and a details button, on the details button you can get more details such as who else is invited to the lunch. And on the B prototype we have a tinder-like interface that lets the user swipe or click accept or decline buttons to search for his or her lunch groups. Here additionally the people that are pending are also visible at first sight and you can go back if you by accident declined a group. The key difference between these two interfaces is the selection process. In the A version we have a listing and in the B version we have one after each other.

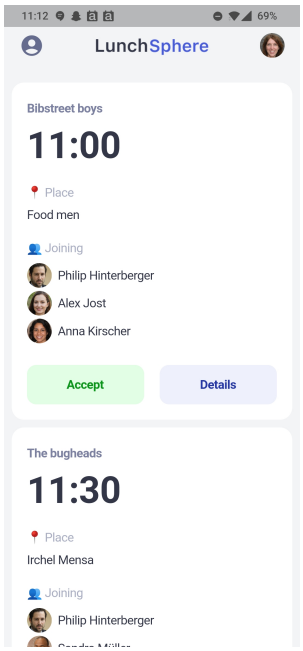


Figure 1: Lunch scheduling A version

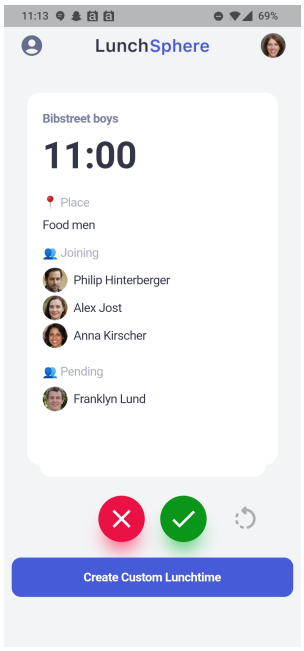


Figure 2: Lunch scheduling B version

## 2 Study Design

- Study Setting: We put the apps on our smartphones (android or ios) for the users to test them in the most authentic way possible. The studies were all conducted one on one at different locations by each member of our group. The study protocol was executed while the data was collected either by the app itself (task execution time, amount of interactions) or by post task questions. After the experiment was conducted we put the data into a google sheets file to enable easy collaboration and easy further data processing.
- Independent Variable The independent variable in this experiment is the lunch scheduling process. The independent variable is either A or B, the A version of the app is a list based interface (see Fig 1.) and the B version is a swiping (Tinder inspired) interface (see Fig 2.)
- Dependent Variable We measured four different Dependent Variables in our test, namely:
  - **SUS Value**: A simple ten question questionaire for assessing the subjective system usability
  - **Interaction Count**: How many times the participant put the finger from the screen
  - **Button Presses**: How many buttons were pressed during the testing.
  - **Task Completion Time** Time passed from opening the app to commiting to a lunch group.
- Hypotheses
  - **NH1**: The swiping feature instead of the button accept feature has no impact on the SUS value of the user
  - **NH2**: The swiping feature instead of the button accept feature has no impact on the time needed to select a group
  - **NH3**: The swiping feature instead of the button accept feature has no impact on the amount of interactions (i.e., taps/swipes) the user has to do.
- Experimental Procedure The experimental procedure in summary looked like this:
  - Introduction to the Experiment
  - Training Phase: Let the participant test and figure out the app for approximately 1 minute.
  - Give the participant our imaginary scenario during lunchtime and then give them the phone to schedule a lunch based on that scenario. It contains time constraints, preferable cafeteria and people you don't want to eat lunch with.
  - Give the participant the SUS rating scheme to fill out
  - Do a post task interview with 4 questions
  - Switch prototypes
  - At the very end ask the participant to compare the two versions of the app and let them express their preferences.
- The participant demographics and summary of the pre-study questionaire is summarized in Figure 3.

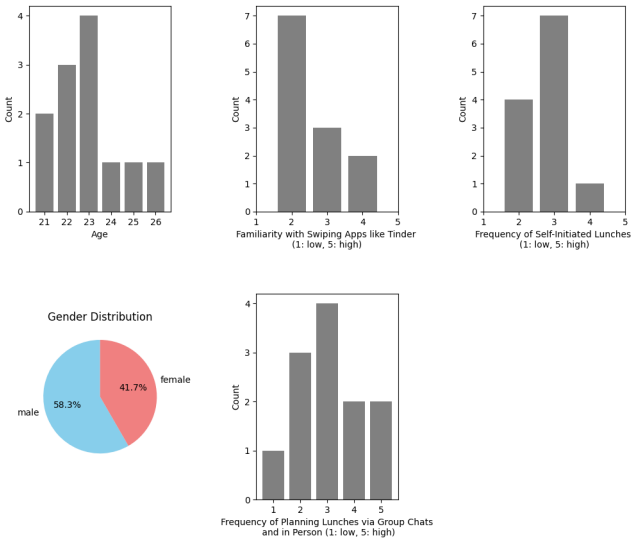


Figure 3: Visualization of participant demographics and pre-questionaire results. From top left to bottom right: age, familiarity with swiping apps like Tinder (1-5), frequency of self initiated lunches (1-5), gender distribution and frequency of making lunches in group chats (1-5)

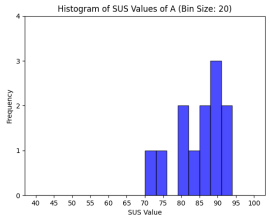


Figure 4: SUS values A version

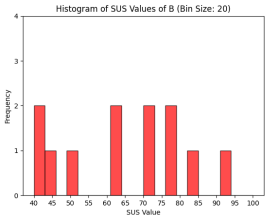


Figure 5: SUS values B version

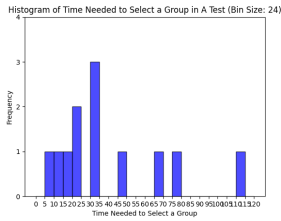


Figure 6: Task completion time A version

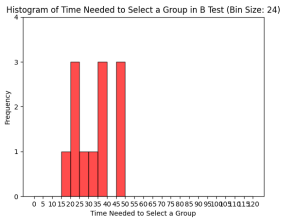


Figure 7: Task completion time B version

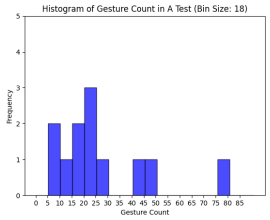


Figure 8: Gesture count histogram A version

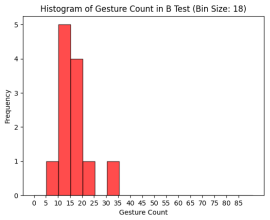


Figure 9: Gesture count histogram B version

## 3 Results

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

Quantitative Results To compare the effect of the scheduling process on the SUS value, we conducted a Wilcoxon Signed Rank Test (all Shapiro-wilk $p>0.05$ and Levene's $p<0.05$). With interface A, participants on average gave a SUS rating of 84.6 (SD = 7.21). With interface B, participants on average gave a SUS rating of 64.5 (SD = 19.1). The mean difference between the two groups was statistically significant; $Z = 7.0$, $p = 0.009$ These results indicate that interface A was percieved as more usable than interface B. NH1 is rejected. Looking at the qualitative results the reason for this might be that a Tinder like interface does not have a good enough overview of all the possible dates and people in general want to see all the options before they commit to a lunchtime with a group

To compare the effect of the interface 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 39.8 seconds (SD = 30.3 sec) to complete the task. With interface B, participants on average needed 32.7 seconds (SD = 10.4 sec) to complete the task. The mean difference between the two groups was not statistically significant; $t(11) = 0.83$, $p = 0.43$ (NEED DISCUSSION HERE???)

To compare the effect of the interface on gesture count, we conducted a Wilcoxon Signed Rank Test (Shapiro-wilk on A $p<0.05$ and $p>0.05$ for B Levene's $p>0.05$). With interface A, participants on average needed 29 gestures (SD = 20 gestures) to complete the task. With interface B, participants on average needed 17 gestures (SD = 6 gestures) to complete the task. The mean difference between the two groups was not statistically significant; $Z = 13.5$, $p = 0.052$ Given this result, we do not have enough statistical evidence to confidently reject the null hypothesis assuming a significance level of 5 percent. H3 is not rejected. (BRAUCHT BEGRÜNUNG HIER?)

Summary of Qualitative Results We asked the participants four questions after each prototype about thoughts, confusion, what they liked or disliked, issues and unexpected behaviours. The most mentioned points for prototype A were:

- Pending status. Most participants did not like that the pending is not shown in the overview.
- Most participants liked the overall design
- Some participants would have wished for better filtering to find the perfect lunch match.

For the prototype B the most mentioned feedback points were:

- The most frequently mentioned aspect was that there is no real overview of all lunches and they wished there was one
- Even though there was a back button present participants said that it is not obvious enough that there is one present.
- Many stated that the swiping is fun and the app looked visually pleasing.

At the end we asked which prototype they liked better and why. The result from that is pretty clear. Every participant liked the A version better. The most frequent reasons were the better overview and the ease of use in the A version. In the B version there was not such a strong consensus. But rather some voices saying that the decline button there and the better overview of the pending status were appreciated.

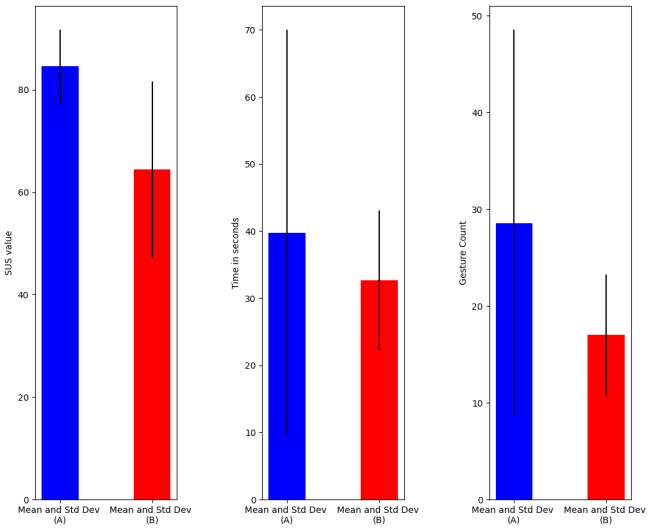


Figure 10: Mean and standard deviation from left to right for SUS value, task completion time and gesture count

## 4 Discussion

The study gave some meaningful insights in our A/B test. The quantitative data gave us a statistically significant difference in the System Usability Score (SUS) between the A and B interfaces. (M = 84.6, SD = 7.21) compared to interface B (M = 64.5, SD = 19.1), Z = 7.0, p = 0.009. This suggests that the users found the list based interface subjectively more usable than the swiping interface. This is in strong parallel with the results form the qualitative data that states that users wanted to have an overview of possible lunch arrangements before commiting to one.
In contrast, the task completion time and gesture count data did not differ enough to make a statistical impact. This might be because the swiping feature in the B version is a bit more unnatural and unusual. People needed more time and gestures with that, while not considering at all the options and often accepting the first best option made the process quicker. The list based A version the interface was directly more familiar and easy to use but there were many more options present in one sight that potentially resulted in more thinking and decision making time.

## 5 Limitations

Limitations in sample size and diversity might limit the generalizability of this study. Additionally the scenario that was given to the participants was not varied and might not have been optimal to test the app versions well enough in all aspects.

## 6 Future Work

Correcting some more peripheral features like the pending status that was criticized, one could redo the study with a larger and more diverse sample size. Having a larger sample size would possibly give more insight about the task completion time and gesture count differences in both versions. As the overview was nonexistent in the B version, one could modify the B version to include some kind of overview while keeping the swiping feature.

## 7 Conclusion

In conclusion, our study gave a well rounded picture in the A/B testing of two interfaces. Having both quantitative and qualitative data we gained a multi perspective view on the matter and are now able to make the very informed decision that the prototype A will be our final one.