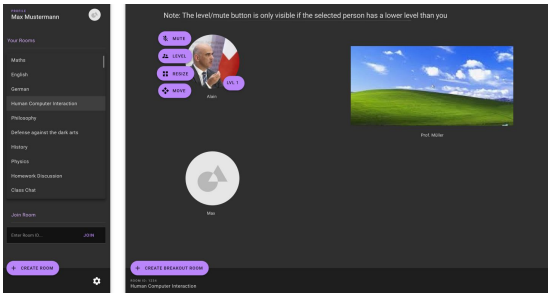
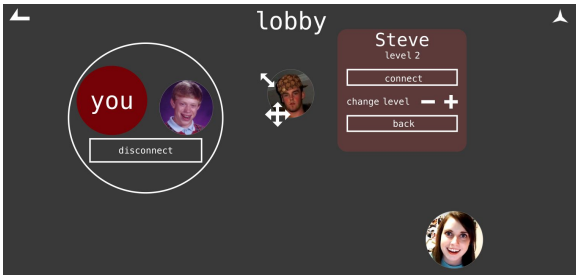
Study Report

Introduction

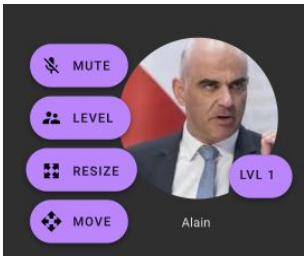
Both interfaces aim to implement the ideas of **user-specific conference arrangement** and **level-based permission system**. Major differences mostly include how and where these additional functionalities can be accessed.

Furthermore, the prototypes propose different approaches for the complexity of the interface. While *Prototype A* maintains a classic sidebar menu for the management of different meetings (similarly to *Teams*), *Prototype B* tries to hide these details and focuses on the ongoing call (similarly to *Zoom*).





Prototype A



Prototype B

Goals

During the study, we pursued the following main goals:

- **How easy is it for the participant to make use of the additional functionalities?**
- **Which prototype is less overwhelming?**

Recall that the initial purpose of this entire project was to get video calls closer to physical meetings by giving participants the ability to somehow naturally form groups without making it too complex or playful. That is exactly what our additional functionalities try to achieve.

While it is rather difficult to test how the user perceives an actual video call using these simple prototypes, it is very feasible to measure his performance and discover whether he experiences these features as beneficial or disturbing.

Study methodology

In order to collect as much data as possible in such a short time by also considering the current situation where physical meetings should be avoided, we tried to make participation as easy as possible for the user by making both prototypes as well as the rating form directly accessible via browser. This allowed the participant to sit at his own device where he usually conducts video calls, led to a very comfortable situation and also had the benefit of eliminating any independent variables regarding the environment.

After a short introduction of all the necessary details regarding the project, the participant was confronted with the first prototype (this could be either *Prototype A* or *Prototype B*) and was given the opportunity to make himself familiar with its functionalities.

Then, the following 4 basic tasks had to be performed while the **task completion time** was measured as a dependent variable:

- "*Create a room and let other people join.*"
- "*Change another user's level.*"
- "*Change the position and size of another user.*"
- "*Create a breakout-room, join it and share your screen.*"

Afterwards, the participant was asked to fill out the rating form for this prototype before the previous steps were repeated with the other prototype.

Person ID	0	1	2	3	4	5	6	7	8	9	10
Age	21	20	55	24	23	21	53	51	22	14	32

Since we compared users of different age groups and computational competence, there were lots of independent variables regarding each participant's background. This however does not influence our results in any negative way as our target group also consists of all these different user types and we are interested in overall average performance and experience.

Results

To compare the effect of the prototype on task completion time, we performed a **paired t-test** analysis. We started with the null hypothesis that none of the prototypes is faster than the other. In order to accept or reject the null hypothesis, we need to check if the averages are reliably different from each other. Note that we search for a lower mean value of each task in order to optimize the user experience. The calculation has been performed with JASP and the results look as follows:

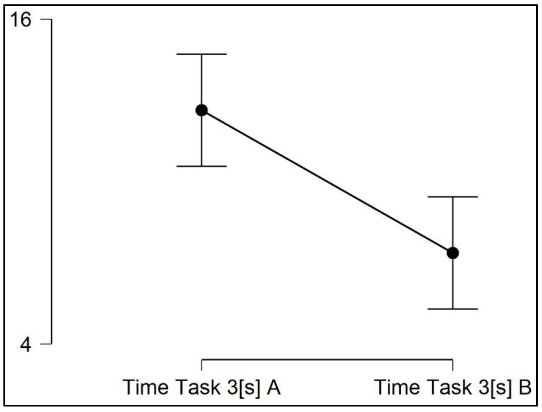
Paired Samples T-Test

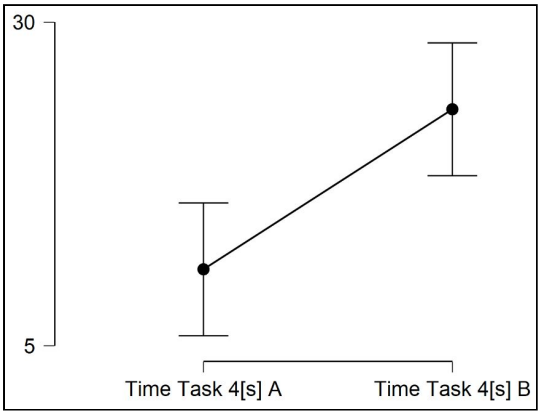
Measure 1		Measure 2	t	df	p	Cohen's d
Time Task 1[s] A	-	Time Task 1[s] B	0.302	10	0.769	0.091
Time Task 2[s] A	-	Time Task 2[s] B	−0.218	10	0.832	−0.066
Time Task 3[s] A	-	Time Task 3[s] B	4.010	10	0.002	1.209
Time Task 4[s] A	-	Time Task 4[s] B	−3.797	10	0.004	−1.145

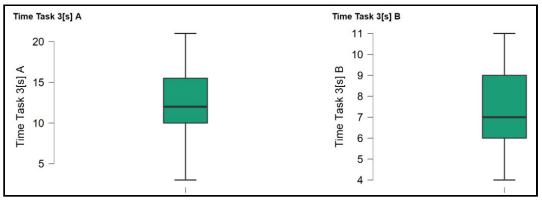
Descriptive Statistics

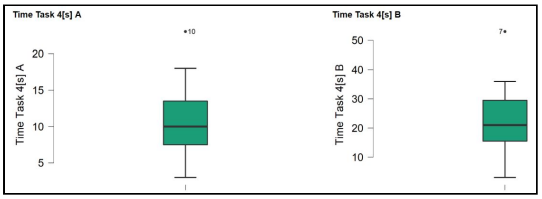
	Time Task 1[s] A	Time Task 1[s] B	Time Task 2[s] A	Time Task 2[s] B	Time Task 3[s] A	Time Task 3[s] B	Time Task 4[s] A	Time Task 4[s] B
Valid	11	11	11	11	11	11	11	11
Missing	0	0	0	0	0	0	0	0
Mean	11.182	10.273	7.727	8.091	12.636	7.364	10.909	23.273
Std. Deviation	6.882	7.630	4.338	3.780	5.104	2.292	5.941	13.425
Minimum	5.000	3.000	3.000	3.000	3.000	4.000	3.000	3.000
Maximum	30.000	31.000	17.000	14.000	21.000	11.000	23.000	53.000

We define the significance level α as 0.05. The difference in measurements can be assumed to be significant if $p \leq \alpha$. Since the p-values of *Task 1* and *Task 2* are way above the significance level, we cannot reject the null hypothesis and no real conclusion can be drawn on both task comparisons. If the null hypothesis applies, no difference between two measured phenomena can be concluded. However, for *Task 3* and *Task 4* the null hypothesis can be rejected. The mean value of the time measured in *Task 3* is higher for *Prototype A*, thus favoring *Prototype B*. The mean value of time measured in *Task 4* is higher for *Prototype B*, thus favouring *Prototype A*.

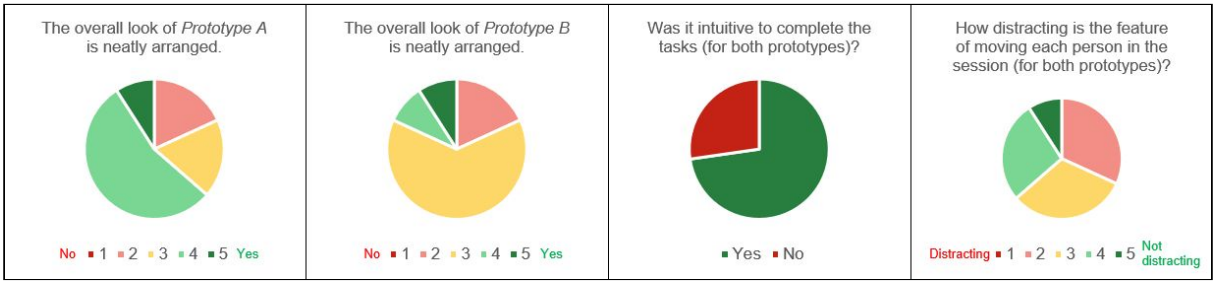








Regarding the subjective rating as well as the verbal feedback we received, *Prototype A* clearly takes the lead over *Prototype B*. Furthermore, the participants generally had a neutral to good experience when completing the tasks. Both of these insights are based on the following diagrams:



We conclude the presentation of results with a quote of *Participant#10*:
"*That's exactly what's missing in Skype for Business! Being able to place the faces by yourself. This allows me to make groups of people that belong together and gives me a better overview.*"

Limitations

All results need to be analyzed with care. 11 degrees of freedom (respective participants) are not enough to get meaningful results. The comparisons are meant as an easy indication of which interface is faster to use and which should be favoured. Moreover, the demographic of the participants are not that well distributed, especially the age.

Conclusions

The study shows only partial consistent results for a particular interface. *Task 1* and *Task 2* show no favourable interface, so we assume both interface designs as equally well. *Prototype B* is favourable in *Task 3*. *Prototype A* is favourable in *Task 4*. We however were indeed able to collect some qualitative verbal feedback regarding our ideas and prototypes. It turns out that the design of *Prototype A* is clearly more popular across the participants.