THE POWER OF CONVERSATION FLOW IN VIDEO CONFERENCE TOOLS: EVALUATION OF SPEAKER CHANGE CUES

Mincheng Chang[1], *Thilo Michael*[1], *Sebastian Möller*[1], *David Schlangen*[2]

[1]*Quality & Usability Lab, Technische Universität Berlin, Germany,*
[2]*Computational Linguistics, University of Potsdam, Germany*
mincheng.chang@campus.tu-berlin.de

Abstract: This project aims at exploring whether the predictive turn-taking model could enhance the quality of conversation flow in video conference tools and the overall user experience of conversations in digital scenarios. We designed a subjective experiment in which participants watched modified video materials that simulate different turn-taking indicator behaviors (predictive model, reactive model, and baseline model) in various scenarios (different amount of speakers, different window layout types, different speaker indication methods) with two conversation contents. The participants are acting as non-contributing over-hearers and score the performance of the system and their opinion towards the dialogue. Post-experiment interviews were being conducted to gather detailed opinions and possible explanations for observed phenomena. The results show that the predictive turn-taking model has a positive significant difference in video conferences and is preferred by most of the participants.

1 Introduction

Due to the massive spreading of COVID-19, the majority of people have to study or work remotely through video conference platforms. Present video conference platforms provide images, audio, screen sharing functionalities so that participants can see each other, understand each other, and accomplish their tasks. However, these solutions are not ideal. The lack of eye contact, conversation flow, physical surroundings, and latency of the internet creates an invisible gap between participants and decreases the overall user experience comparing to face to face conversations. Nevertheless, individual designers and commercial companies are working on solutions that can enhance the overall user experience from different aspects of video calls. Individual inventor *Matt*[1] creates a reflective and physical installation with a reflective mirror, transparent panel, and a camera on top of the laptop's screen to regain eye contact between participants. In widely-used online video conference platforms *Zoom*[2] and *VooV*[3], they provide some digital solutions, for example, the function of adjusting the color of videos, changing background into pictures or videos and even beautifying the portrait of speakers. In contrast to all of the solutions mentioned above, which are mainly focused on image quality and visual effects, we are more interested in the conversation flow. Pre-experiment interviews indicate that the clear speaker change cues might be one of the reasons for people preferring face-to-face conversations. People have the feeling of control of the conversation or are capable to follow the conversation when they know who is the next speaker in advance.

[1] `https://www.youtube.com/watch?v=2AecAXinars`
[2] `https://zoom.us/`
[3] `https://voovmeeting.com/`

2 Related Works

Research in Kousidis & Schlangen (2015) [1] has shown that a robot that follows a conversation is perceived favorably when it uses a predictive turn-taking model, which means anticipating speaker changes and potentially facing it's head to the next speaker before the end of the speaker's utterance, compared to when it uses a reactive model, which means moving it's head only after the end of the utterance. The observed effect could be powered by a modern artificial intelligent algorithm and could have dramatic potential in improving user experience. Existing research projects are aiming at achieving the predictive algorithm for multi-view video coding with a different focus, such as focusing on content-aware prediction algorithm with inter-view mode decision [2] or the efficient prediction structures [3]. These projects are focusing in the aspect of data transmission, signal processing, and machine learning. There are problems such as longer latency due to higher demand of computing and data transmission or relatively high error rate. Instead of working on an iteration of the existing prediction algorithm, we focus on the user's perception of the conversation flow and we investigate potential indicators, which enable smooth turn-taking from human behavior. Research shows humans tend to predict the next speaker by noticing turn-yielding signals and attempt-suppressing signals from present speakers in the face-to-face conversations [4].

3 Experiment Design

Because it is difficult to control human behavior to precisely achieve certain signals in an experiment environment. We decided to use existing videos of video conference calls and edit the turn-taking indicator around the video frame to simulated the desired effects. This makes it more feasible to research the conversation flow in limited conversation contents which have arranged speaker sequences rather than discussions or casual chatting. This paper will explore the effects of different turn-taking indication strategies in modern standard video conference scenarios with clear speaker change cues, by referencing tools such as Zoom, in which the next speaker is usually reactively indicated by a green frame, a bigger window, or both.

We created 18 simulated video call stimuli that consist of 2 conversation scenarios, each with three simulated speaker turn-taking models and three layouts. The three turn-taking models are a predictive model, where the indication changes before the next speaker's turn, a reactive model, where the indication changes after the speaker's turn, and a random model, that changed the indication at random pre-defined intervals. The three video layouts were a layout with 6 participants, where one participant is shown in a larger video in the center, as well as a video with 6 participants and with 16 participants, where the videos are aligned in a grid.

In the experiment, the participants are acting as non-contributing over-hearers in the video conference videos. We use the within-subject method to conduct the experiment, which means each participant has to watch 18 videos including every scenario, and answer 36 questions in total. To balance the learning effect, we use the Latin square method [5][6] to create 18 sequences of video orders and participants will randomly choose one sequence. After each video, the participants answered two questions: "How well do you believe the system could follow the conversation (by indicating the next speaker)?" and "How close was the behavior of the system (track next speaker) similar to what you expect a human do in face to face conversation?" for each video we presented. The first question is designed to ask about objective opinions of the system's intelligence and the second question is designed to ask about subjective opinions of overall user experience compared to physical conversations.

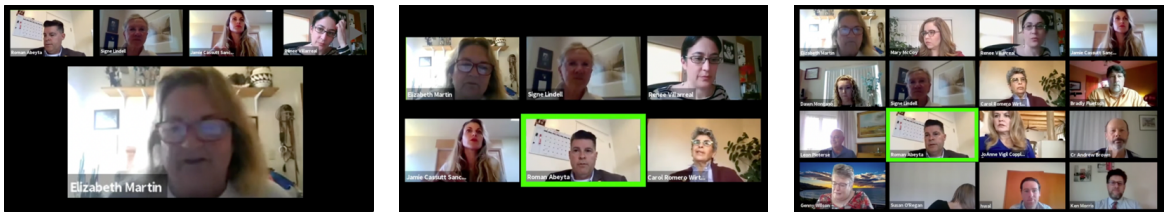
3.1 Experiment Materials

3.1.1 Source Videos

We use public video conference recordings from *YouTube*[4], the *Source Video 1*[5] is about several vote call process and the *Source Video 2*[6] is about an interview with football team. To create a scene with 16 participants, we also included non-contributing persons from *Source Video 3*[7]. The conversations in these videos are in English.

3.1.2 Layouts & Indication Methods

According to the analysis of Zoom's system behavior models, there are different windows layout models such as speaker model, grid model, list model, and so on. In terms of speaker indication methods, the Zoom uses a green frame as a baseline method to emphasize the speaker and a big window as an extra method in the speaker model. To control variables, we use the green frame method in the grid and only the big window method in the speaker model. According to pre-experiment interviews, we discovered that the participant numbers might have a strong influence on part of interviewees' feelings in a video conference. We decided to set different participant numbers in the video materials. There are originally 4 combinations of layout and indication methods. Due to a limited number of windows in the speaker model, there is no chance for the no-contributing persons to show up. So we choose only 3 combinations as shown in the figure 1, namely 6 participants, where the active speaker is shown in a big window, 6 participants aligned in a grid, and 16 participants aligned in a grid.



(a) 6 Participants * Big Window **(b)** 6 Participants * Green Frame **(c)** 16 Participants * Green Frame

Figure 1 – Layouts and Indication Methods from Source Video 1

3.1.3 Turn-Taking Models

We have designed three different turn-taking indication methods for the selected source videos: a predictive model where the next speaker is anticipated, a reactive model where the speaker is indicated shortly after they started speaking, and a baseline model that pseudo-randomly selects a participant. We edited the source video to simulate the desired effect (see figure 2). We keep the audio track from the original video and we modified the timing of different indication methods such as the green frame and the bigger window. To simulate the real scenarios in the physical world, we keep a 0.3s time difference as internet delay and add a 0.3s time difference as speaker's reaction delay in the reaction model. For the prediction model, we switch the indication methods when the previous speaker stops speaking and sometimes indicate the

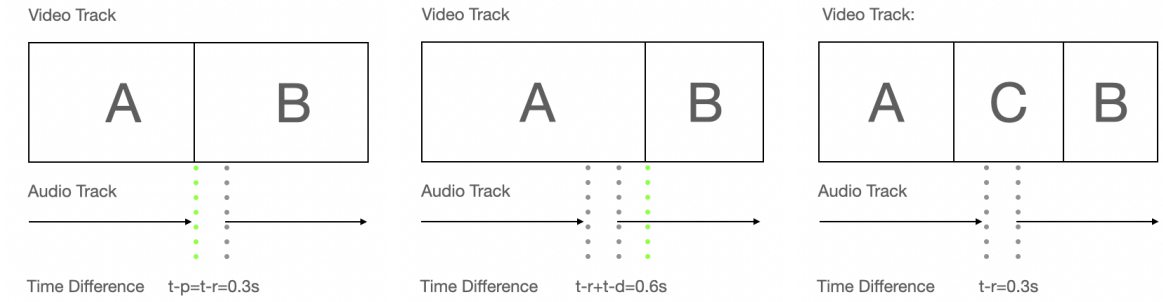
[4] https://www.youtube.com/

[5] https://www.youtube.com/watch?v=Fa-hbpZx6OQ&list=PLgItt8nUkcqYvh86SM-ZtBE4KKuYvQ2BX&index=13&t=167s/

[6] https://www.youtube.com/watch?v=GfTzVjeSuxs&list=PLgItt8nUkcqYvh86SM-ZtBE4KKuYvQ2BX&index=19

[7] https://www.youtube.com/watch?v=60jEQcJUgpI&list=PLgItt8nUkcqYvh86SM-ZtBE4KKuYvQ2BX&index=15&t=433s

next 3 speakers to simulate the effect of an optimal prediction algorithm. The baseline model is designed to simulate the worst case that for every speaker turning, the indication method for each turn-taking is decided by a randomly generated number sequence. The indication methods in the baseline model include predictive, reactive turn-taking models and indicating wrong speakers.



(a) Predictive Model (Optimal) **(b)** Reactive Model (Reality) **(c)** Baseline Model (Random)

Figure 2 – Design and Realization of 3 Turn-Taking Models

3.2 Questionnaire

For the experiment we used TheFragebogen [7], an online HTML5 frame work. The questionnaire includes three parts: a demographic questionnaire and 18 videos, each with two questions in form of a NASA-TLX rating scale [8]. Participants were required to watch the whole video and answer all questions to continue to the next video.

3.3 Data Analysis

We recruited 20 participants (12 male and 8 female), aging from 19 to 28, coming from universities and companies. The participants all speak English as a second language. They have used video conference tools for study or work for at least one year. Thus, no additional introduction to video conferencing was necessary for this experiment.

We conducted statistic analysis with scatter plots, spearman signed-rank correlation analysis, box-plots, and the post-hoc test method. The NASA-TLX rating scale has an advantage in relatively representing participants' opinions. Participants could select his or her range for scoring and if there is a very good/bad feeling, they could score at a position outside of their original range. Considering that participants have different standards and different ranges for scoring, we decided to focus on the scope of changes. We normalized the participant's data between 0 and 1 separately with the minimum and maximum scores of his/her answers. This way, the highest score represents the most preferred settings and the lowest score represents the most unfavorable settings.

4 Result

4.1 Correlation between user ratings

A spearman signed-rank correlation was used to test if question 1 (the perceived intelligence of the system) and question 2 (the user experience) correlate (see figure 3). There was a moderate to high positive correlation between the two variables($r = 0.6792$, $p \ll 0.05$). We could state that the perceived higher intelligence of systems can lead to an overall higher user experience.

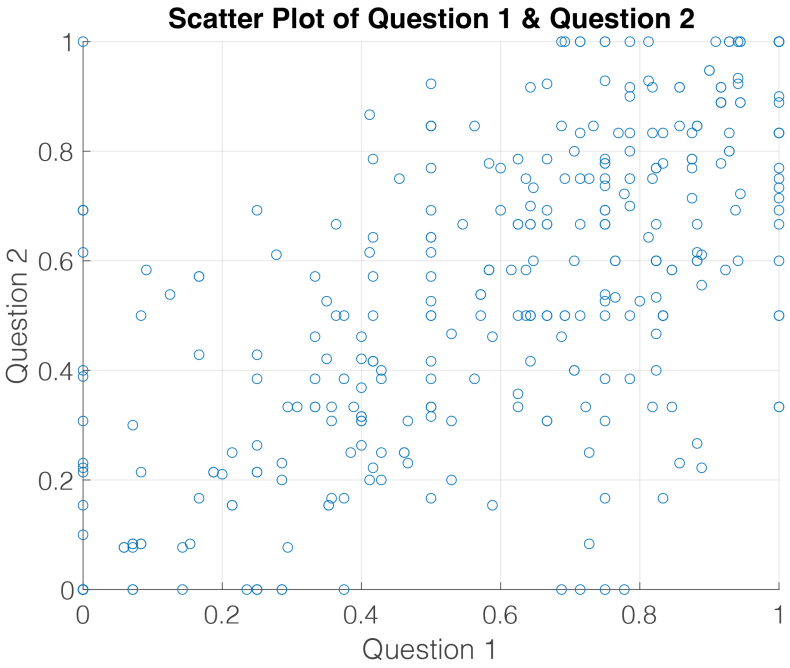


Figure 3 – Scatter plot of ratings (question 1 vs question 2)

4.2 Box Plots & Ranks of Rating-General

The distributions of the ranks of ratings for question 2 is shown in figure 4. The distributions of the ranks of ratings for question 1 is identical to question 2. The One-way ANOVA test revealed that there are statistically significant difference in turn-taking models between at least two groups ($F = 8.99$, $p \ll 0.05$ for question 1; $F = 7.59$, $p \ll 0.05$ for question 2). The post-hoc test shows that the predictive and reactive turn-taking models are significantly better than the baseline model in both question 1 and question 2.

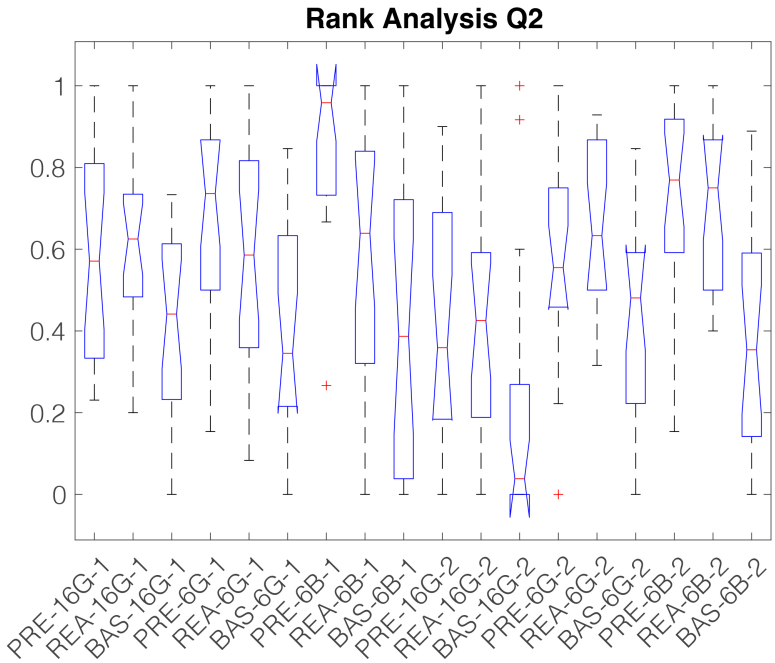
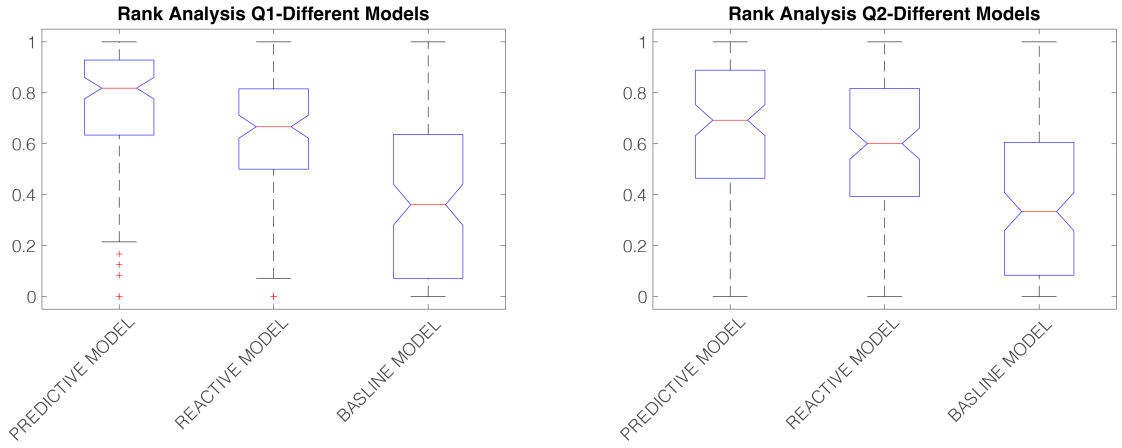


Figure 4 – Box Plots of Ranks of Rating for Question 2. Lines show the medians and squares show the means

4.3 Box Plots & Ranks of Rating-Models

The distributions of the ranks of ratings for questions 1 and 2 in the category of different turn-taking models are shown in figure 5 respectively. The One-way ANOVA test revealed that there are statistically significant difference in turn-taking models between at least two groups ($F = 58.08$, $p \ll 0.05$ for question 1; $F = 32.76$, $p \ll 0.05$ for question 2). The post-hoc test shows that the predictive turn-taking model is significantly better than the reactive turn-taking model and baseline model.

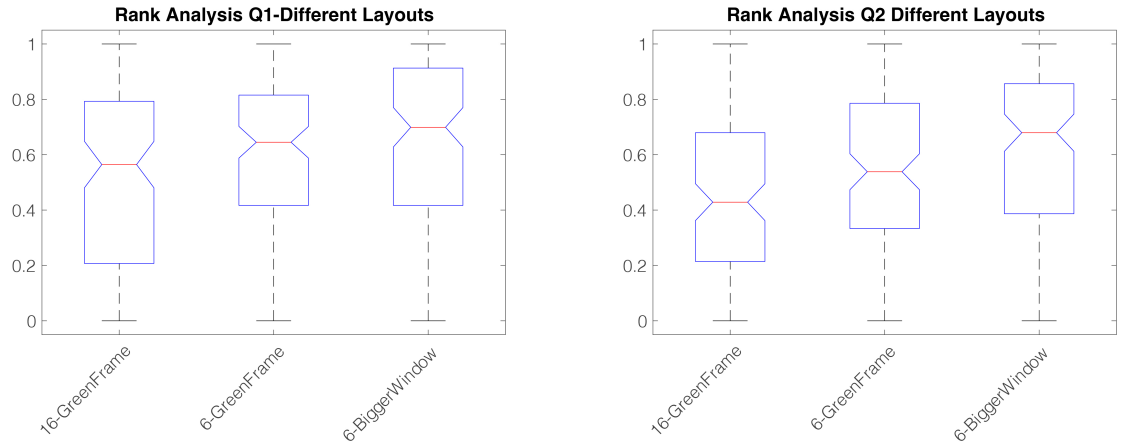


(a) Question 1-Models **(b)** Question 2-Models

Figure 5 – Box Plots of Ranks of Rating for Question 1 & 2 in Category of Different Models. Lines show the medians and squares show the means

4.4 Box Plots & Ranks of Rating-Layouts

The distributions of the ranks of ratings for questions 1 and 2 in the category of different layouts are shown in figure 6 respectively. The One-way ANOVA test revealed that there are statistically significant difference in turn-taking models between at least two groups ($F = 6.9$, $p = 0.0012$ for question 1; $F = 9.74$, $p \ll 0.05$ for question 2). The post-hoc test shows that the 6-window-bigger-window layout is significantly better than the 16-window-green-frame layout in question 1 and question 2, even when we analyze the data in categories of predictive turn-taking model and reactive turn-taking models.



(a) Question 1-Layouts **(b)** Question 2-Layouts

Figure 6 – Box Plots of Ranks of Rating for Question 1 & 2 in Category of Different Layouts. Lines show the medians and squares show the means

4.5 Ranks of Rating-Videos

According to the ranks of rating, we observed no significant difference between video 1 and video 2 for question 1 ($F = 1.41, p = 0.2365$), and a significant difference for question 2 ($F = 4.09, p = 0.0438$). However, the effect size r ($r = 0.3410$) for question 2 is medium, and the Cohen's f ($f = 0.1316$) for question 2 is small. Thus, we consider there is no fundamental difference between video 1 and video 2 for both questions.

5 Discussion

The result shows participants are generally holding a positive attitude towards the predictive turn-taking model in video conference tools. We found some counter-examples when we did data analysis for every single participant's scores. Combined with our post-experiment interviews, we got participants' opinions and reasons behind these facts.

While the predictive turn-taking model has on average higher ratings, the reaction model gets higher scores from some participants. According to the post-experiment interviews, these participants think the system should react to the behavior of the speaker which is a more natural and preferable way for them. This part of participants might be used to the present video conference tools and they prefer the tools which are similar to their preference of turn-taking indication. For this phenomenon, we are considering adding additional video materials which shows a conference in the physical world to recall the memory of smooth change of speakers.

Secondly, some participants didn't find a strong difference between the prediction model and reaction model and they scored both turn-taking models similarly and in various small ranges. The reason is that the next speaker in the predictive turn-taking model is indicated only approximately 0.6s faster than in the reaction turn-taking model. The designed 0.6s time difference, which is noticeable for the majority of the participants, yields no difference in rating for them due to individual perception difference. In the future, we aim at researching an adaptive algorithm for calculating the time difference to create an optimal user experience for every single user.

Lastly, participants rated video 1 slightly higher than video 2 in question 2 (for user experience). The reason is that the time of turn-taking in video 1 is 26 and in video 2 is 6. Participants' overall positive feelings are enhanced by the higher time of turn-taking. The post-experiment interviews show that participants have the feeling of mastering the conversation for both video 1 and video 2. They have concrete expectations for the process of vote call and interviews which have clear speaker change cues. When the name of the next speaker is called by the present speaker, the participants would start researching him/her in the windows especially in the grid model. Despite the behaviors of speakers, participants tend to focus on the names speaker mentioned, the intonations of questions, and the pacing of the ending. These 3 indicators might be valuable for constructing the algorithm for predictive turn-taking model in the future.

Nevertheless, we are researching about what is the optimal solution for video conference tools shortly. The predictive algorithm might provide some insights and possibilities. However, this research only proves that the predictive turn-taking model is valuable in video conferences with clear speaker change cues such as vote calls and interviews. When it comes to other conversation contents such as daily chatting or discussions, more complicated theories or effects of optimal predictive turn-taking model are needed to be researched, designed, and tested. Peo-

ple's feelings and standards change from time to time when they adapt to the digital world or are used to present video conference tools. Further follow-up research is also needed.

6 Conclusion

We have presented in this paper the current state of our ongoing research. We have made 18 video materials to simulate the effects of predictive, reactive, and baseline turn-taking models in 9 scenarios. We have conducted experimental research with 20 participants and analyzed the experimental data. We could state that the predictive turn-taking model is perceived favorably and it can enhance the overall user experience in modern video conference tools. In addition, we figured out that the predictive turn-taking model has a more significant influence in the 6-window-bigger-window layout setting and it is usable and suitable in scenarios with traceable conversation contents such as vote calls or interviews.

In the future, we plan to design and make a video conference prototype that either integrates a predictive turn-taking algorithm or provides functions that enable participants to actively indicate the next speaker or speakers. And we will continue researching with experiments in live conversation scenarios.

References

[1] KOUSIDIS, S. and D. SCHLANGEN: *The power of a glance: evaluating embodiment and turn-tracking strategies of an active robotic overhearer*. In *2015 AAAI Spring Symposium Series*. 2015.

[2] DING, L.-F., P.-K. TSUNG, S.-Y. CHIEN, W.-Y. CHEN, and L.-G. CHEN: *Content-aware prediction algorithm with inter-view mode decision for multiview video coding*. *IEEE Transactions on Multimedia*, 10(8), pp. 1553–1564, 2008. doi:10.1109/TMM.2008.2007314.

[3] MERKLE, P., A. SMOLIC, K. MULLER, and T. WIEGAND: *Efficient prediction structures for multiview video coding*. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11), pp. 1461–1473, 2007. doi:10.1109/TCSVT.2007.903665.

[4] DUNCAN, S.: *Some signals and rules for taking speaking turns in conversations. Journal of personality and social psychology*, 23(2), p. 283, 1972.

[5] FISHER, R. A.: *Statistical methods for research workers*. In *Breakthroughs in statistics*, pp. 66–70. Springer, 1992.

[6] PREECE, D.: *L atin squares, l atin cubes, l atin rectangles, etc. Encyclopedia of statistical sciences*, 2004.

[7] GUSE, D., H. R. OREFICE, G. REIMERS, and O. HOHLFELD: *Thefragebogen: A web browser-based questionnaire framework for scientific research.* In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019.

[8] HART, S. G. and L. E. STAVELAND: *Development of nasa-tlx (task load index): Results of empirical and theoretical research.* In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.