Toward Optimal Field-of-View for Captioning on Head Worn Displays

Asheton Arnold
Georgia Institute of Technology
Atlanta, Georgia
Aarnold40@gatech.edu

Laird Stewart
Georgia Institute of Technology
Atlanta, Georgia
Laird.Stewart@gatech.edu

Omar Joudeh
Georgia Institute of Technology
Atlanta, Georgia
Ojoudeh3@gatech.edu

Brittney Bush
Georgia Institute of Technology
Atlanta, Georgia
Brittney.Bush@gatech.edu

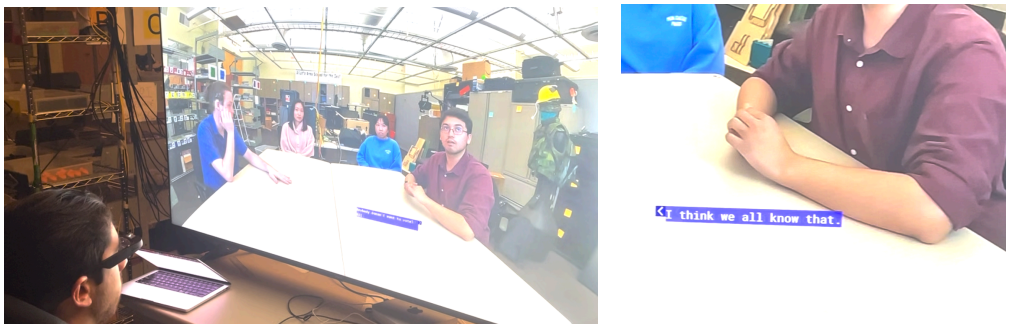


Figure 1: Georgia Tech's Contextual Computing Group's HWD Simulator

Abstract

One useful application for head-worn displays (HWD) is providing live captioning and translation. Live captioning on HWDs would benefit the d/Deaf or hard of hearing (d/DHH) community and be useful in noisy industrial environments. We explore the preferred FOV of captions within the user's visual field in group conversation.

A recent study performed by Britain et al. [1] found that 30-degree non-registered captions are preferred over 10, 20, or 40 degrees. We build on this study by focusing on FOVs of 15, 20, 25, and 30 degrees with a pilot study. Our research participant's stated preference suggests 25 and 30-degree wide captions are generally preferred. 25-degree captions, on average, outperform the others based on NASA's task load index (TLX) indicator, though our sample size (n=12) is too small to provide statistically significant results.

Introduction

430 million people are d/Deaf or hard of hearing (d/DHH) globally, and the WHO projects that 1 in 10 people will be affected by hearing loss in 2050. [2] These individuals can be greatly benefitted from assistive technology, like hearing aids and cochlear implants. These devices are not perfect and have very high costs associated with them and repairs. [3] Another form of assistive technology has come about with improvements in audio processing and mobile technology: live transcription on mobile phones. Head-worn displays (HWD) can be used to make transcribed conversations easier to follow and decrease the visual dispersion between the captions and the speaker. Understanding the minimum and optimal FOV of captions can help manufacturers and designers create functional and discreet HWDs.

We explore the preferred horizontal width of non-registered captions with indicators within the user's visual field in group conversation. Non-registered captions are displayed in the same location of the user's FOV as opposed to registered captions which remain "tied" to the speaker. Indicators point in the direction of the speaker. [1] We conduct research (n=12) within a virtual HWD captioning simulation to hold other HWD variables constant and identify the optimal caption width between 15, 20, 25, and 30 degrees.

Our research participant's stated preference suggests 25 and 30-degree wide captions are generally preferred. 25-degree captions, on average, outperform the others based on

NASA's task load index (TLX) indicator, though our sample size (n=12) is too small to provide statistically significant results.

Previous Work

Georgia Tech's Contextual Computing Group (CCG) has conducted a range of studies on HWD captioning. Jason Tu et al's Towards an Understanding of Real-time Captioning on Head-worn Displays [4] compares the captioning between a HWD and a mobile phone for a toy block construction activity for hearing participants. A majority of the participants claimed they would prefer captioning on a HWD if the voice recognition functioned better, despite the fact that the trial was conducted twice with two distinct HWDs and both times the HWD required more mental effort and caused irritation. Another study by the same authors, Conversational Greeting Detection Using Captioning on Head Worn Displays Versus Smartphones [5], explored whether participants preferred greeting detection on a phone or a HWD; the results indicated that a HWD was preferred by the participants and offered faster greeting detection than a mobile phone.

Also from the CCG, Britain et al's Preferences for Captioning on Emulated Head Worn Displays While in Group Conversation separately compares caption delivery method and width. The study's participants favored registered (captions are placed under the current speaker) and non-registered (the captions follow the user's head movement) captions with indicators (an arrow showing the direction of the speaker) for comfort, ease, speaker identification, and readability. In the second experiment, the authors compared caption widths of 10, 20, 30, and 40 degrees. NASA-TLX and Likert results improved from 10 to 20 to 30 degrees but declined from 30 to 40 [1].

Our study continues the work of Britain et al. and narrows in on the ideal caption width, comparing 15-, 20-, 25-, and 30-degrees wide captions and expands Britain et al's group conversation captioning simulator (Appendix F).

Our Work

Our work consisted of two parts: Improving Britain et al's group conversation captioning simulation and conducting a pilot study comparing caption width.

First, we smoothed caption rendering along the x-axis. In the legacy code, the text would shake sporadically leaving the text almost unreadable. The code filtered the Google Glass' IMU sensor data using a moving average of the last 500 angle measurements. We tested the moving average with different sampling widths and an exponential filtering algorithm. The fundamental tradeoff in sampling widths was between smoothness and lag. Ultimately a simple moving average with a sample width of 20 provided the most readable text.

Next, we streamlined the captioning simulator. We added a command line argument to update the caption width which was previously hard-coded. We also used a headrest and taped measurements to standardize participants' distance from the screen which was previously ignored. This is key to ensure that our calculations are correct for the size of the caption box. These adjustments made our research run more smoothly and accurately.

Finally, we wrote a Python script to calculate caption width in pixels and characters given the simulation setup and FOV angle. We chose a monospace font to ensure precise width. This will help future researchers if they need to change the simulation monitor or focal distance.

Methodology

We ran 12 participants through the captioning simulator. Each was shown the same video of a group conversation four times with different caption widths. Widths were determined using a latin square to remove any bias due to familiarity with the simulation setup or video content. After each video, each participant took the NASA TLX (Appendix B) [6] and answered a few extra questions on a Likert scale regarding their experience. In the end, we asked a few short unstructured interview questions to gain qualitative feedback on the simulation and captioning.

Quantitative Results

25-degree captions slightly outperformed the other widths for each NASA TLX category (Appendix A). 30-degree captions were assessed as second best for most categories while 15-degree and 20-degree captions performed generally worse and roughly the same. However, a one-way ANOVA test found the differences in sample means not to be statistically significant (Appendix C).

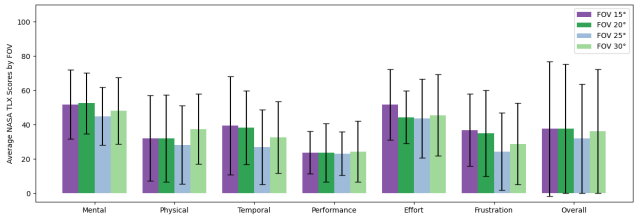


Figure 2: NASA-TLX scores (Larger version in Appendix A)

Averaging the responses to our task-specific, Likert scale questions suggests 25 and 30-degree captions were slightly

more preferred (Appendix D). Like the NASA TLX responses, these differences were marginal.

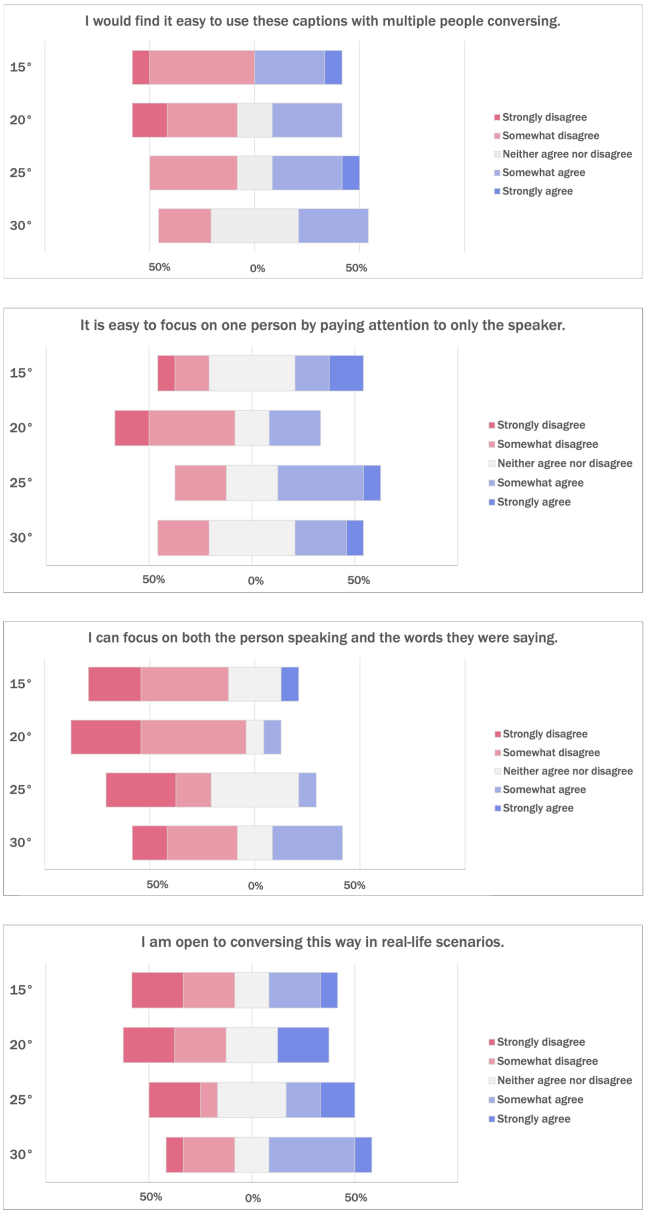


Figure 3: Likert Scale scores (See Larger in Appendix D)

Qualitative Results

Un-structured, qualitative discussions with our participants provide insight into the participant's stated rationale for these responses. Most subjects (n = 5) pointed to looking between captions and faces as a major difficulty in following the conversation. Participants claimed that looking at the actor's faces for queues, expressions, or reactions took their attention away from the captions causing them to miss details. Stated preference towards preferred caption width varied. Some (n = 4) preferred wider captions while others (n = 3) preferred narrower ones. Participants in the former group claimed they preferred wider captions because each word lasted more time on screen, making it easier to keep up. One in the latter group preferred shorter captions because the distance between the actor's faces and the end of each caption was smaller, conducive to easier tracking back and forth.

Discussion

Our team set out to improve the CCG's captioning simulator and run a pilot study on hearing individuals to assess the simulation's performance and refine the study design. We successfully smoothed the captions by filtering the sensor data and provided support for dynamically altering the caption width via the command line. Our pilot study discovered new limitations with the experiment, and flaws with the simulation, and helped refine the procedure.

Our study suggests that on average, 25 and 30-degree non-registered captions with indicators are preferred and improve self-reported TLX scores when compared to 15 and 20-degrees. However, this preference and task performance vary from person to person, and our sample size was not large enough to provide statistically significant results. These results are consistent with Britain et al's [1] findings which suggested 30-degree captions outperformed 10, 20, and 40 degrees (Appendix E). Combined, our studies move towards identifying the optimal width for non-registered captions with indicators in group conversation.

This analysis is consistent with our unstructured, interview feedback. It appears there is a fundamental tradeoff between wider captions increasing the duration of each word and narrower captions keeping words closer to the speaker. In both studies, captions between either extreme, around 25-30 degrees, perform best.

Limitations

We hold a variety of variables constant to isolate changes in caption width. First, the simulated focal length of the display equals the distance to the actors. Captions remain sharp and in focus, assuming perfectly adjusted lenses without eyebox issues. Also, only one actor speaks at a time to avoid dealing with conflicting captions or indicators. Finally, each caption is restricted to two lines of text with the same font style, size, transparency, and color (Appendix F).

The simulation also has room for improvement. First, the caption box's left edge is aligned with the center of the visual field. For longer captions, this caused participants to noticeably correct their view towards the left to center captions on the actors. Additionally, the vertical location of

the captions does not track the participant's head movement, nor was its location in the visual field standardized for all participants.

Further, we assume that our trial subject's backgrounds do not influence their performance or survey results. For example, subjects who often read captions or are more comfortable with HWDs may find the task easier or more enjoyable. We do use a latin square to order our experiments to reduce any learning bias or gained comfort with the system.

Future Work

In the future, we would like to improve the HWD simulator, improve our research procedure, and test new caption widths with the d/DHH community. From our unstructured interviews, we identified a few simulator pain points which distracted from the experience. First, despite our improvements, the lag and shakiness in our caption rendering made it difficult for our participants to read the captions. Occasionally, the consistency of the filter was so unreliable that we had to re-start the simulator. This could be improved by using a better IMU sensor than the Google Glass or improving our filtering algorithm. Next, our captions' left bound is centered on the user's visual field, not the middle of the caption itself. Changing this would be straightforward and would improve our simulations' accuracy. Finally, we found the duration that each captioned phrase lasted on the screen directly impacted our user's stated preference. Therefore, one possible direction of future research would be to make the number of lines each caption is rendered with a variable in the simulator.

As for our research procedure, we identified a learning curve to utilizing the indicators; participants were unsure what and where they were in the beginning. There is a study done by D. Jain et al. that investigates different ways of signaling sound loudness and direction using both egocentric and exocentric models that would be beneficial in creating more obvious and usable indicators [7]. A second learning curve through the process was that the user must move their head for the simulator to render the captions properly, not their eyes. One solution for both issues would be to run a test video for each subject before the experiment to get them acquainted with our setup.

In the future we hope to run our experiment on members of the d/Dhh community as they are the end user and differ from the average hearing participant in that they have more experience reading captions. It appears that the ideal caption width within the user's field of view is around 20- or 25-degree. Therefore, we would like to run an experiment directly comparing the two angles and run a paired, two-way, two-tailed t-test to analyze the results. An a priori analysis with standard values of $\alpha = 0.05$ and $1- \beta = 0.8$ finds a sample size of (n=34) participants to be ideal.

Conclusion

For this project, we continued the work done by the CCG group surrounding COG and set up a study procedure to find the optimal caption width within the visual field needed for effective captioning. We improved legacy code used to simulate captioning on HWD to better the experiences of both the researcher and the participant and afterwards ran a trial study. From that study we saw a preference towards 25 and 30 degrees, although the results were statistically insignificant with our number of participants. The simulation had a reported learning curve, which could explain our results as there were other factors distracting the participants from noticing changes in caption width. Next steps for this project include running this study with target individuals, or members of the d/DHH community.

References

[1] Britain, Martin, Kwok, Sumilong, Starner, "Preferences for Captioning on Emulated HeadWorn Displays While in Group Conversation", Draft, 2022

[2] World Health Organization. 2019. Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[3] Bailey, A. (2022, October 27). Hearing aids: Types, features, prices, reviews, and more. Hearing Tracker. Retrieved December 14, 2022, from https://www.hearingtracker.com/hearing-aids

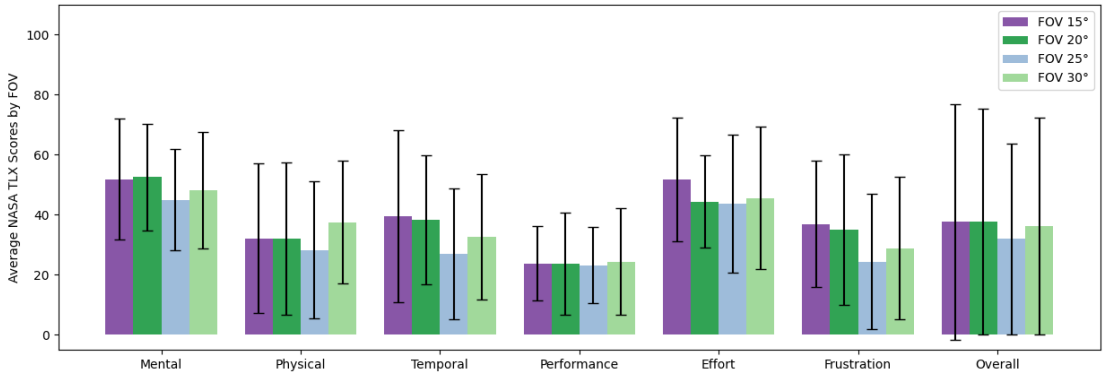
[4] Tu, J., Lin, G., amp; Starner, T. (2021). Towards an Understanding of Real-time Captioning on Head-worn Displays. MobileHCI '20: 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, 1-5. https://doi.org/10.1145/3406324.3410543

[5] Tu, J., Lin, G., amp; Starner, T.(2020). Conversational greeting detection using captioning on head worn displays versus smartphones. ISWC '20: Proceedings of the 2020 International Symposium on Wearable Computers, 84-86. https://doi.org/10.1145/3410531.3414293

[6] So, P., & Gore, B. (2020). TLX @ NASA Ames - Home. NASA. Retrieved November 8, 2022, from https://humansystems.arc.nasa.gov/groups/tlx/index.php

[7] Jain, Dhruv, Leah Findlater, Jamie Gilkeson, Benjamin Holland, Ramani Duraiswami, Dmitry Zotkin, Christian Vogler, and Jon E. Froehlich. "Head-Mounted Display Visualizations to Support Sound Awareness for the Deaf and Hard of Hearing." Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015. https://doi.org/10.1145/2702123.2702393.

Appendix A: Average NASA-TLX Response by FOV



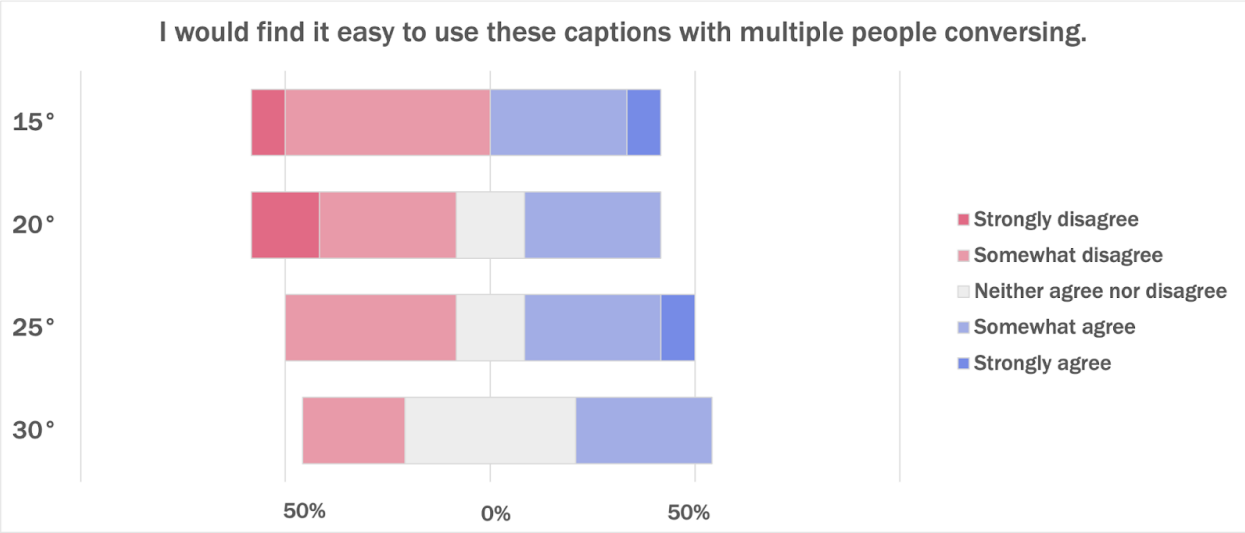
Appendix B: NASA-TLX Questions

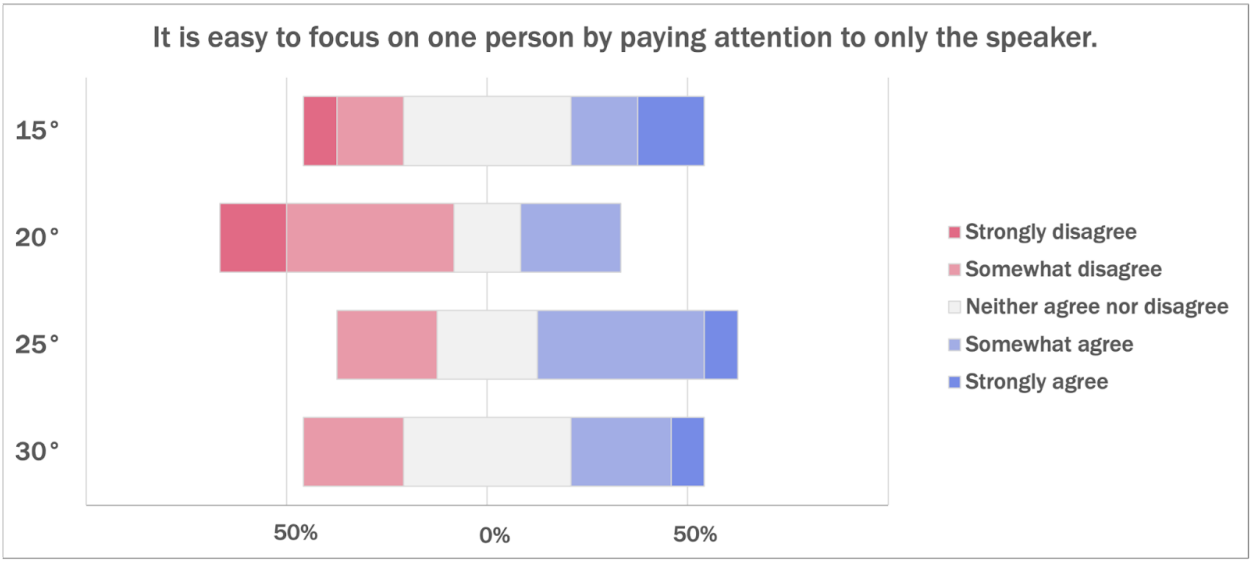
Measurement	Question	100 Point Scale
Mental Demand	How mentally demanding was the task?	Very Low – Very High
Physical Demand	How physically demanding was the task?	Very Low – Very High
Temporal Demand	How hurried or rushed was the pace of the task?	Very Low – Very High
Performance	How successful were you in accomplishing what you were asked to do?	Perfect – Failure
Effort	How hard did you have to work to accomplish your level of performance?	Very Low – Very High
Frustration	How insecure, discouraged, irritated, stressed, and annoyed were you?	Very Low – Very High
Overall	NA	NA

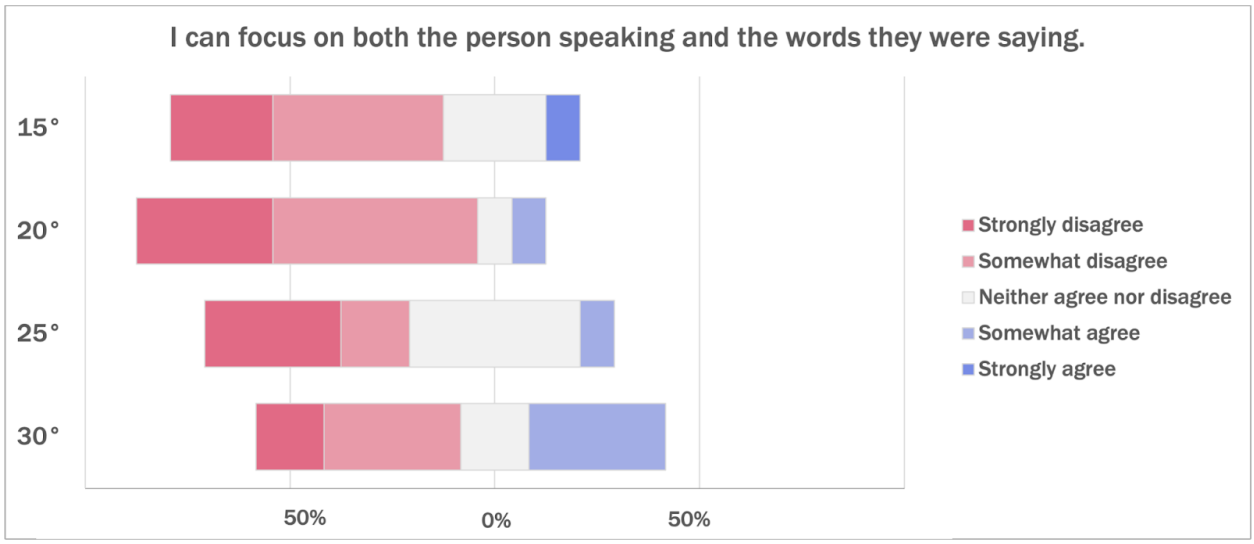
Appendix C: ANOVA Test of NASA-TLX Performance for 15, 20, 25, and 30-degree FOV

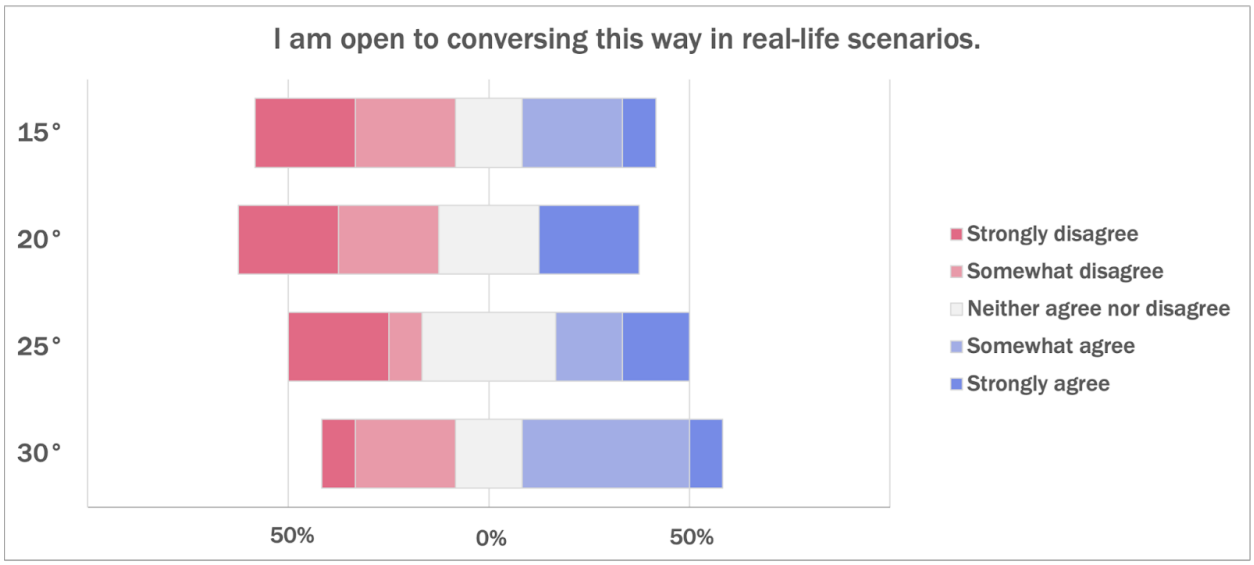
Measurement	F Statistic	p-value
Mental Demand	0.249	0.861
Physical Demand	0.183	0.906
Temporal Demand	0.42	0.739
Performance	0.007	0.999
Effort	0.221	0.88
Frustration	0.43	0.732
Complete Score	0.402	0.752

Appendix D: Average Likert Responses

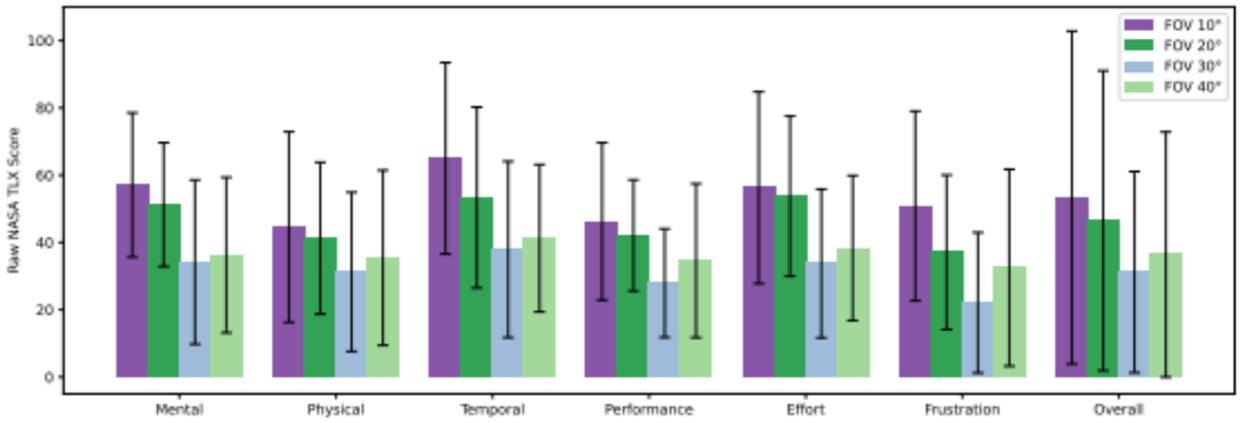





I can focus on both the person speaking and the words they were saying.
15°
20°
25°
30°
50%
0%
50%
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree


I am open to conversing this way in real-life scenarios.
15°
20°
25°
30°
50%
0%
50%
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Appendix E: Britain et al's Average NASA-TLX Response by FOV



Appendix F: Non-registered Captions with Indicators Displayed in the Simulator