WebET 3.0 - Validation Study Report

R&D Team, iMotions A/S

20 Jun 2023

iMotions WebET 3.0 Validation Study

iMotions released WebET 3.0 in May 2023, which is the best webET algorithm in the development of the product to date. The purpose of the validation study was to run a large scale study to evaluate how the algorithm performs on a truly diverse, global, sample, “in-the-wild”. Data was globally collected from 255 participants over 35 days. Participants conducted a short study comprising of gifs, images, videos, and surveys. Self-reported parameters for ethinicity, eye-color, wearing glasses or not, having facial hair or not, and lighting conditions in the room were evaluated against accuracy. Over 50% of participants had an accuracy of 2 degrees of visual angle (dva) or lower. Over 70% had an accuracy of 3dva or lower and over 90% had an accuracy of 5dva or lower. Of the parameters measured, only the presence of glasses had a significant effect on accuracy. The individual differences for ethinicities, regions, eye-colour and the presence of facial hair did not have a significant impact as people collected data in their natural environments Over time, fixation classification stays stable in the center of the screen but classification may reduce in accuracy towards the bottom corners of the screen. Longer studies and internet problems can cause problems with participant compliance and a suboptimal user experience. Therefore, researchers are advised to keep studies as short as possible in order to ensure high levels of participant compliance and optimal data quality throughout the study.

1. Rationale of the validation study

iMotions released a new and improved webET algorithm WebET 3.0 in May 2023. While this is the best webET algorithm in the development of the product to date, with an accuracy of under 3dva under ideal circumstances(Click to read the whitepaper).The purpose of the validation study is to inform clients surrounding the best practices with the usage of this webcam eye tracking in real life situations. The whitepaper evaluates, in detail, which factors can influence webcam based eyetracking. A large scale study such as this sheds light on what a representative sample can look like for researchers and which of the factors isolated in the whitepaper dissipate in a large sample and which ones are more likely to amplify.

Questions asked

1. **What is the accuracy distribution of a dataset collected with WebET 3.0?**
   iMotions has a few ways of ensuring compliance, good data quality, and reporting values that help researchers determine which datasets are good enough to be included in the study. The study looks at how the accuracies are distributed, and if the recommended controls are followed.

2. **Do individual and demographic variables affect accuracy?**
   As people working with eye tracking know, individual and demographic factors such as ethnicities, color of the eyes, people wearing glasses or not, facial hair can all influence accuracy of eye tracking data. We wanted to check to what extent these factors have an influence on eye tracking data in our WebET 3.0.
3. **How much of an impact does lighting have on accuracy?**
   When people have the flexibility of collecting data in the comfort of their homes, they are not able to control the environment as well as a lab study would. Assuming people are seated correctly in front of the camera, the biggest problem here is the lighting condition people complete the study in. The study therefore also asked participants to evaluate the lighting condition they are in.
4. **How does accuracy change with time?**
   Two factors can influence the length of a study. The first is the degree of compliance of participants to sit still for longer periods of time if the study is designed to be very long. The second is internet issues that may create lags in video presentations creating a much longer study for some participants than others. The study aimed to evaluate if either of these factors particularly influenced the accuracy scores.
5. **How does the accuracy translate to using AOIs?**
   Finally, since most people would like to analyze the various kinds of studies and understand what their accuracy scores mean for their research, we look at AOIs taking 5% of the screen, distributed across different locations on the screen and over the length of the study.

2. Methods

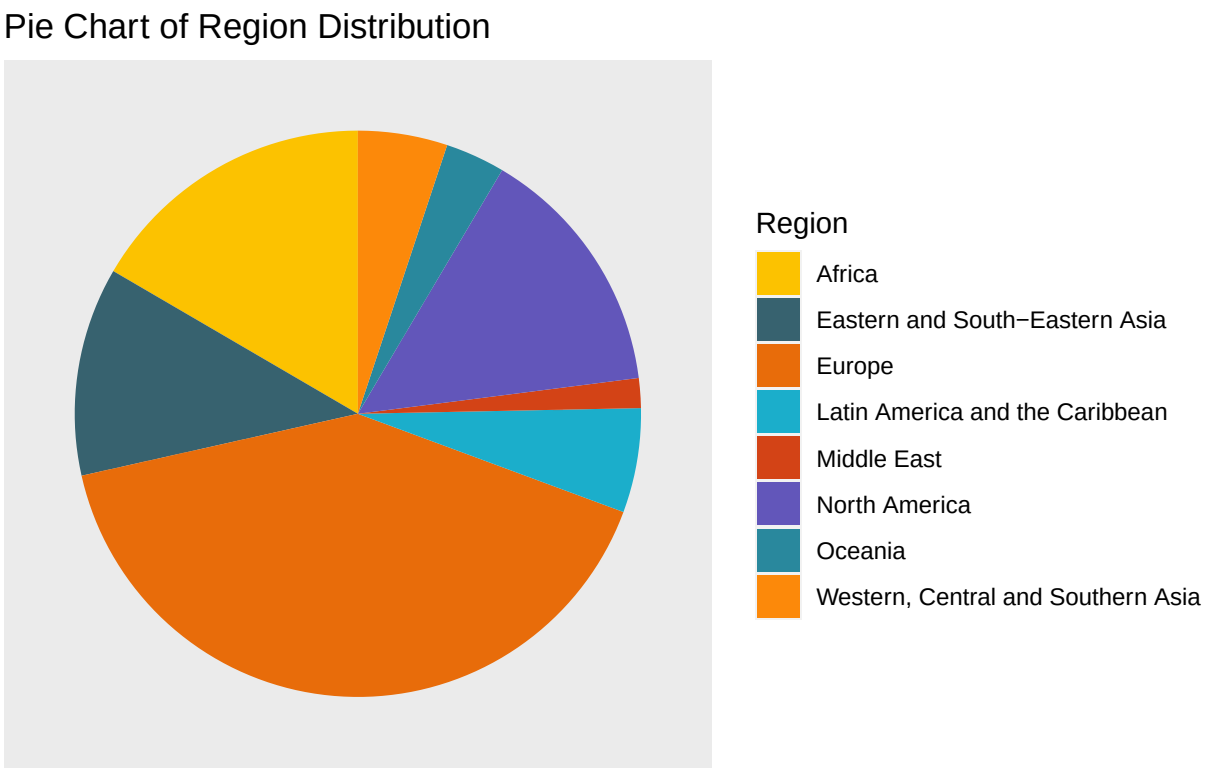
We designed a task lasting less than 10 minutes comprising surveys, gifs, images and videos. The length of study as well as the combination of stimuli is typical of what most iMotions clients currently use.

Study Design The study started and ended with 13 point calibration with inter-stimuli calibration points presented periodically within the study. No restrictions were set on the camera resolution. Participants could use any camera they had connected to their computers.

After calibration, participants were asked to answer a number of survey questions. In the survey slides participants were asked about demographic and environmental conditions (Questions 2 and 3). Following the survey was a block of cat gifs and emojis, used as validation points (Question 5). Cat gifs at 9 points were used to direct people's attention, followed by emojis at the same locations respectively, which acted as the validation targets. The locations were randomized to control for any temporal effects over the 9 positions. These locations did not overlap with the locations of the calibration crosses. This block was repeated before the post-calibration at the end, to evaluate how the answer to Question 5 changes over the course of the study. Between the validation blocks, three animal videos were chosen to introduce some jitter in the length of the study (Question 4) across internet connections.

Participants Participants were recruited from the iMotions community as well as Prolific to ensure a good representation across ethinicities, gender, and eye colour. The iMotions community was contacted via a newsletter explaining the study, and Prolific was used as the panel provider to

target different demographics. The data collection lasted 35 days. At the end of the 35 day period, the study had 255 participants spread out across the world. The geographical distribution of the participants is shown in the below pie chart.



3. Analysis

The analysis was aimed at answering each of the five questions laid out in section 1 and does this one-by-one in the following sections.

Terminology

Before we proceed, here are some helpful guidelines on how to understand the terms being used

- Accuracy: refers to estimated degrees of the visual angle, as calculated by the webET algorithm. This is the score experimenters can see per participant on their online platform. Lower the number, better the accuracy.
- Histograms: are visualizations showing the frequency (on the y-axis) of the continuous event, split into bins (on the x-axis).
- Box plots: For the box plots in this report, each yellow point is a participant with the group membership on x-axis. The boxes show the mean and standard deviation of the group distribution.
- Fixations: After the gaze points have been calculated by the WebET algorithm, fixations were classified using the Hidden Markov Models (I-HMM) algorithm in iMotions. Details of the

HMM are given below.

- For the Area Of Interest (AOI) Analysis, AOIs which were 5% of the screen size were drawn over the cat emojis (validation points) shown.
- Kruskal-Wallis test: Non-parametric test to check if two or more groups are significantly different from each other.

HMM calculation iMotions' Hidden Markov Model is implemented using the HiddenMarkov R package (RStudio Team, 2022). In line with Salvucci & Goldberg (2000), the model initially assumes a probability of 0.95 to stay within each state (fixation or saccade) and of 0.05 to transition to the other state. To start out, the model is given 15 °/s as average gaze velocity during a fixation with a standard deviation of 15 °/s, and 60 °/s as average velocity during a saccade with a standard deviation of 30 °/s. These parameters are then optimized using a Baum-Welch algorithm in order to find the unknown parameters of the Hidden Markov Model. A Viterbi algorithm then applies the results of this optimization and classifies whether a sample belonged to a fixation or saccade.

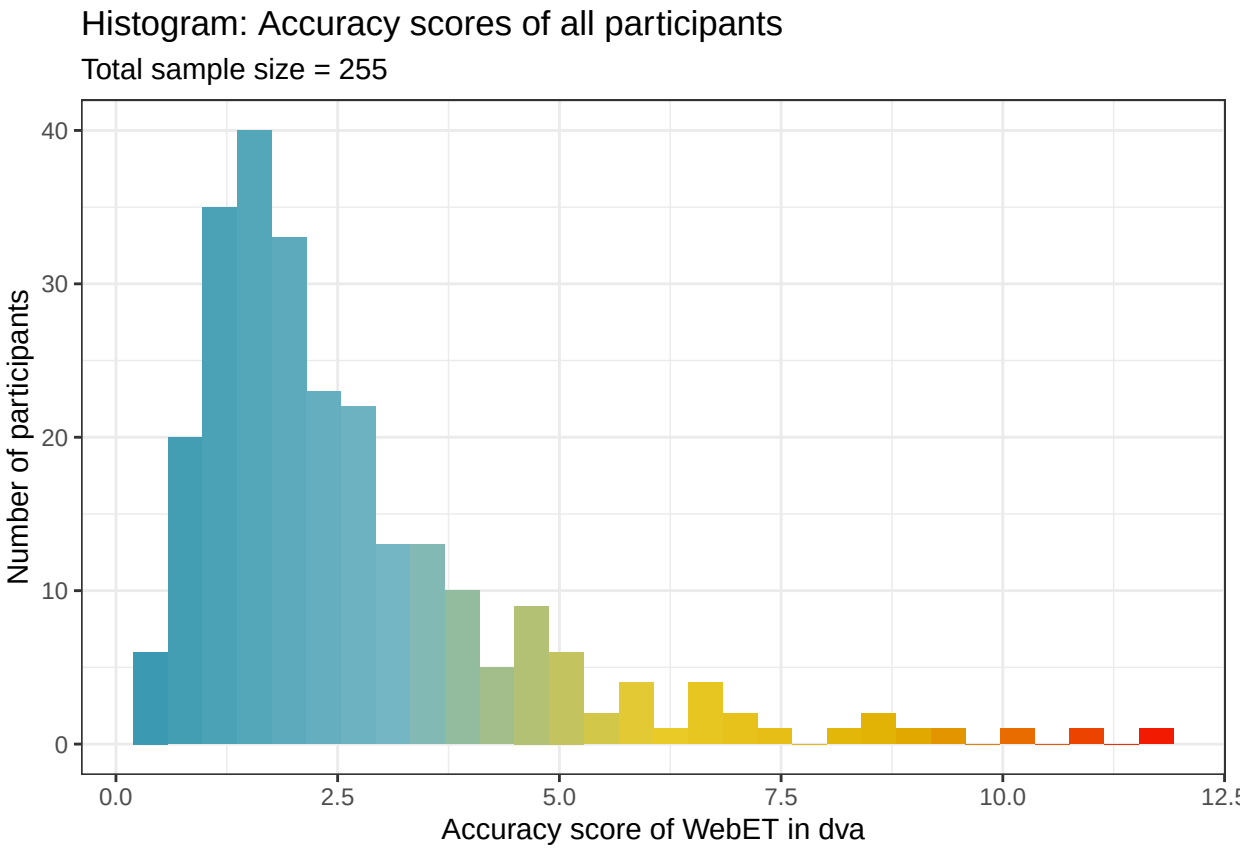
Corrective Steps: Fixations shorter than 60 ms were discarded. Max time between fixations was set to 75ms and max angle to 0.5 degrees, so that adjacent fixations could be merged.

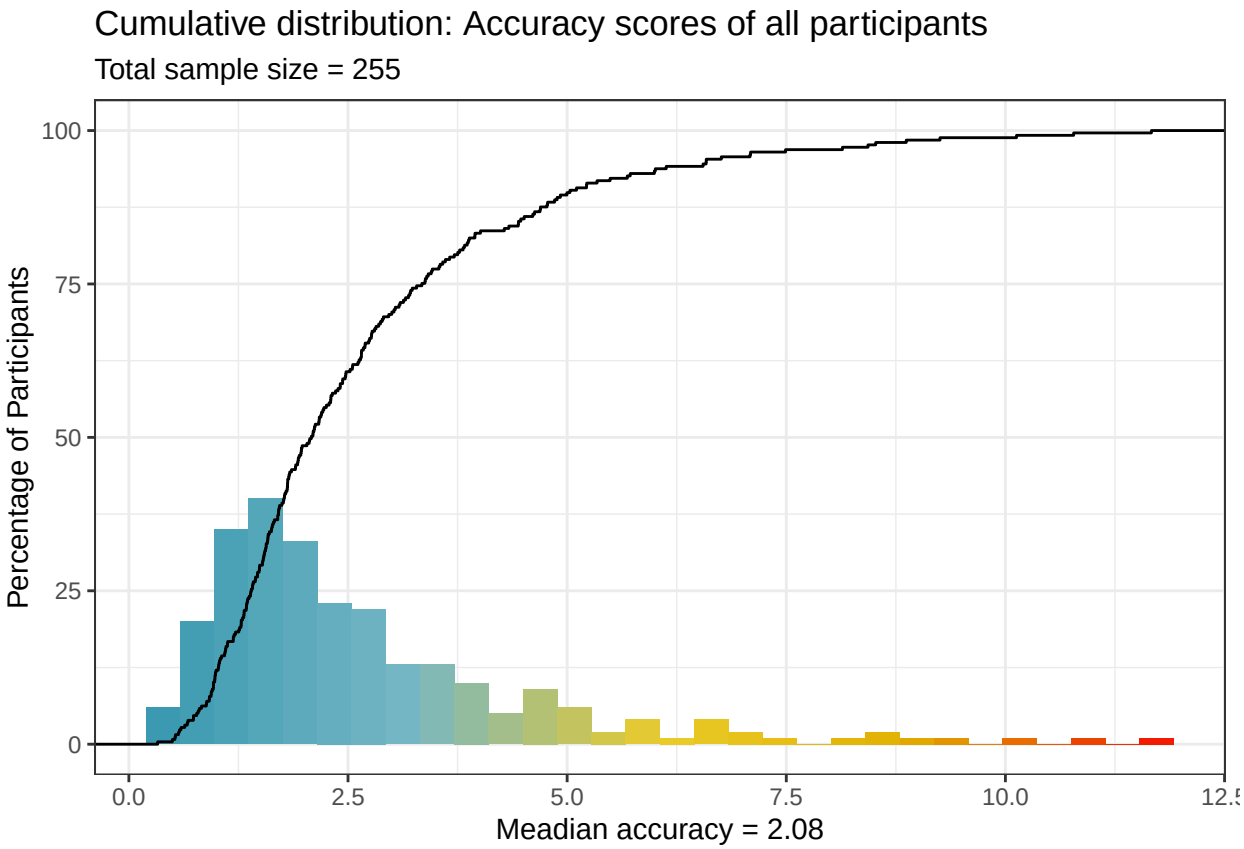
Calculation of fixations: Fixations were numbered from start of the recording to its end, as well as from start of the stimulus to its end. Only fixations with their start and end on the same stimulus get counted as belonging to this stimulus. The start time of the fixation was calculated as the average between the timestamp of the first sample of the fixation and the sample preceding it. The end time of the fixation was the average of the timestamps of the last sample of the fixation and the one succeeding it. The time difference between fixations' start and end times is equal to the fixation's duration. The x- and y coordinates of the fixation's centroid were determined by finding the point with the closest distance to all samples of the fixation.

Methodological Considerations for WebET The field of Webcam-based eye tracking does not have the same history of well-established methodologies compared to infrared-based eye tracking. It is therefore important to keep some considerations in mind while reading the following sections. First, accuracy, as calculated from the calibration slides is measured in dva. While we asked for participant's screen size, there is still no way to track the participant's distance from the screen. An assumption that fits everyone is bound to lead to some unknown offsets.
Second, the AOI analysis in Question 5 looks at areas marked on the validation images, i.e. the cat emojis. The size of the AOI is relative to the stimuli size. This was ascertained at study design and every individual participant, on their varying screen-sizes was resized to fit the stimulus size so percentage of AOI in pixels can be ascertained irrespective of individual screen differences in data collection.
Third, the demographic and environment variables were self-reported. The researchers did not go through every individual recording and measure how well-lit the room and face were. This could introduce biases as participants may not have been able to judge lighting conditions with complete consistency.

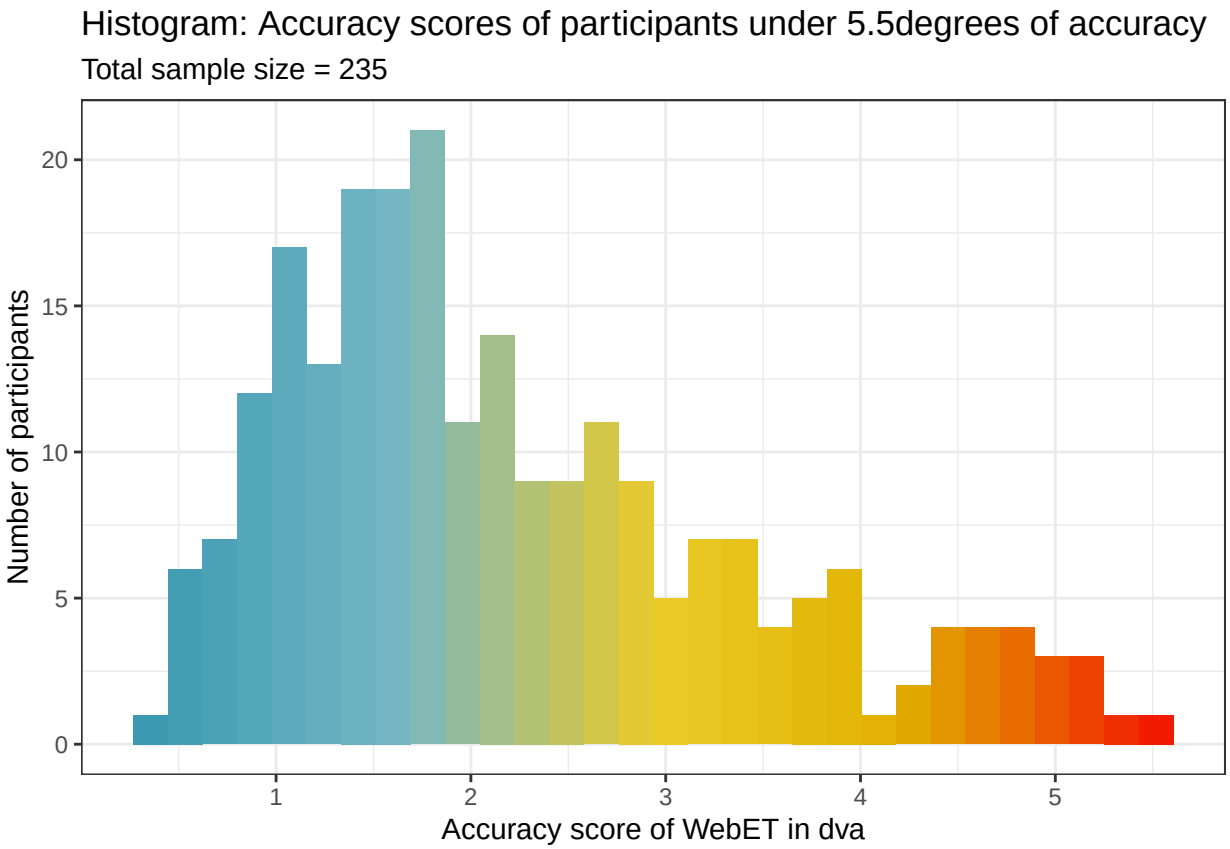
Question 1 : What is the accuracy distribution of a dataset collected with WebET 3.0? At the end of data collection, the validation study had 255 processed datasets. The first histogram below shows the accuracy of all 255 participants.

The same histogram overlaid with cumulative distribution shows 92% of these participants (N = 235) had data below 5.5 degrees of accuracy, and 70% (179) participants had below 3 degrees of accuracy. The median accuracy score was 2.08.



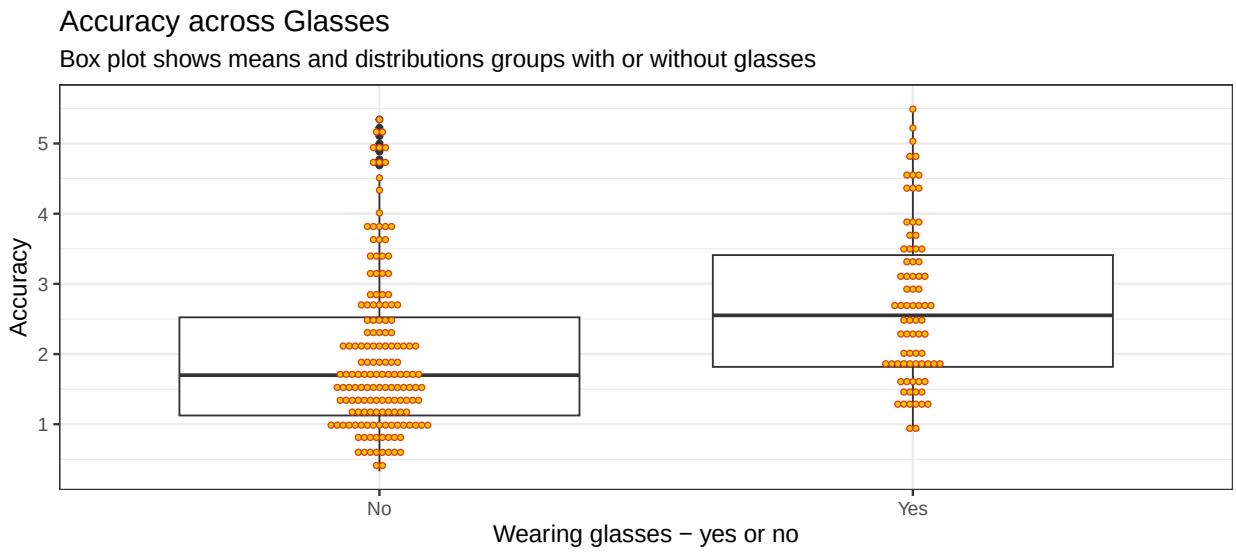


The next histogram is of participants below 5.5 degrees of accuracy. These participants were included in the next steps of analysis.

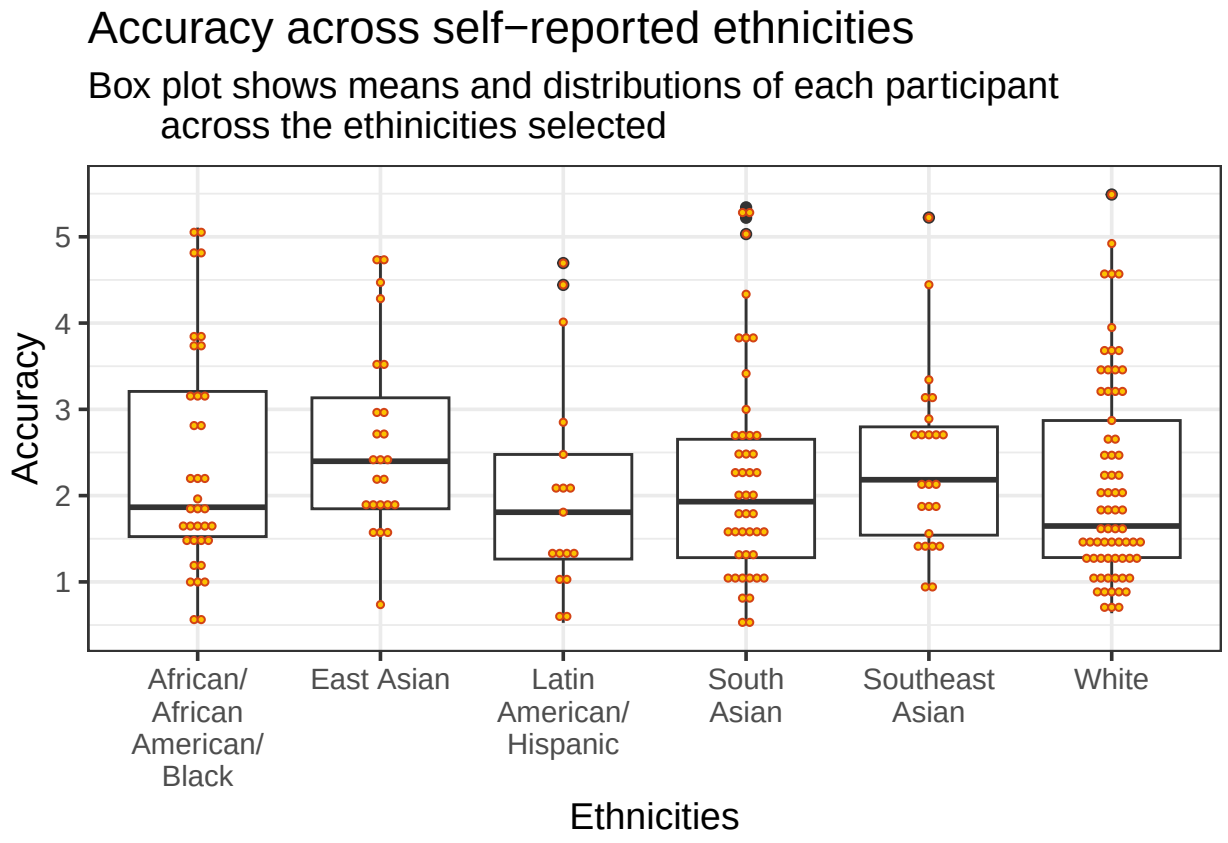


Question 2 : Do individual and demographic variables affect accuracy? The demographic variables that were evaluated to look for group differences on accuracy were - ethnicities, color of the eye, wearing glasses, having facial hair, gender and age. Only one participant reported their gender as “fluid/non-binary”, and less than 5 participants reported having amber or gray eyes, less than 5 participants were under the age of 18 or over the age of 65. These participants were all removed from the respective analysis.

The box plots below show a large variance of accuracy within each factor. While all accuracy differences could not be attributed to any one of these reasons, wearing glasses or not seemed to have a significant difference ($H(1) = 22.56$, $P < 0.0001$) on accuracy scores. This follows the findings of the whitepaper with the mean and median values for people with glasses being higher than those of people without glasses.

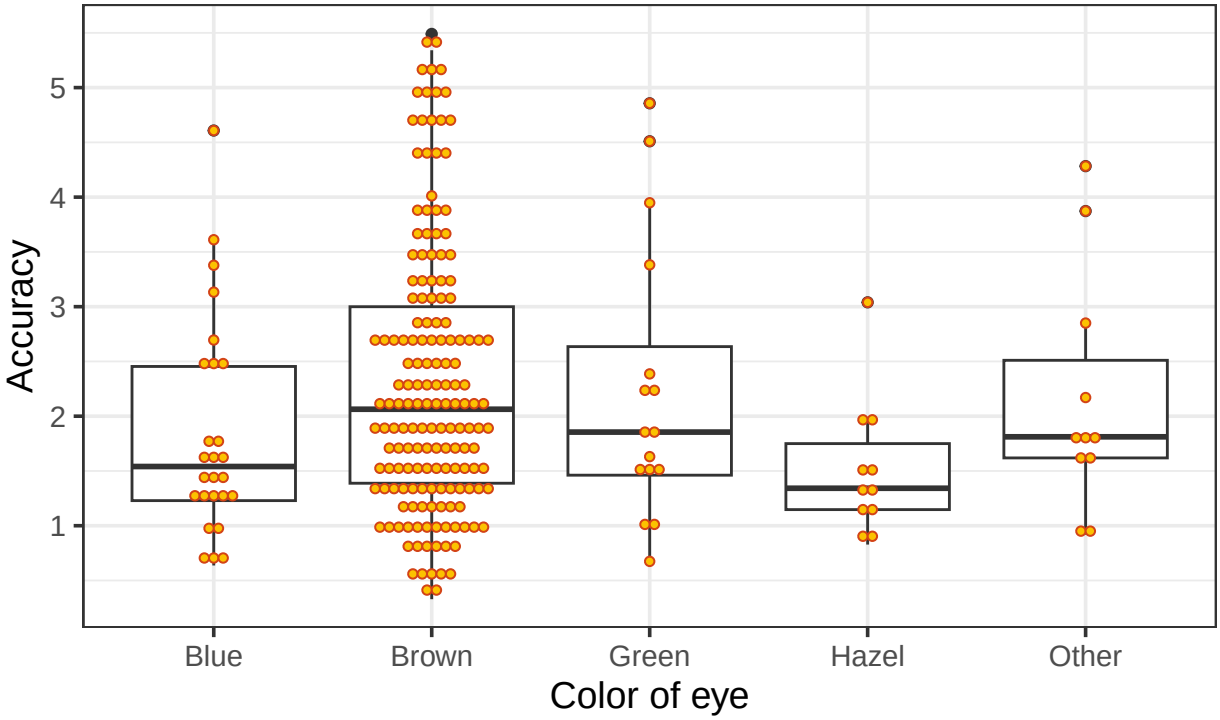


As for the other variables, a study with larger samples such as the present one had no significant differences owing to ethnicities of participants (H (5) = 7.66, P = .18), the color of the eye (H (4) = 8.15, P = .09), if they reported having facial hair or not (H (1) = 1.62, P = .20) their gender (H (2) = 3.69, P = .15) or age (H (6) = 8.84, P = .18).



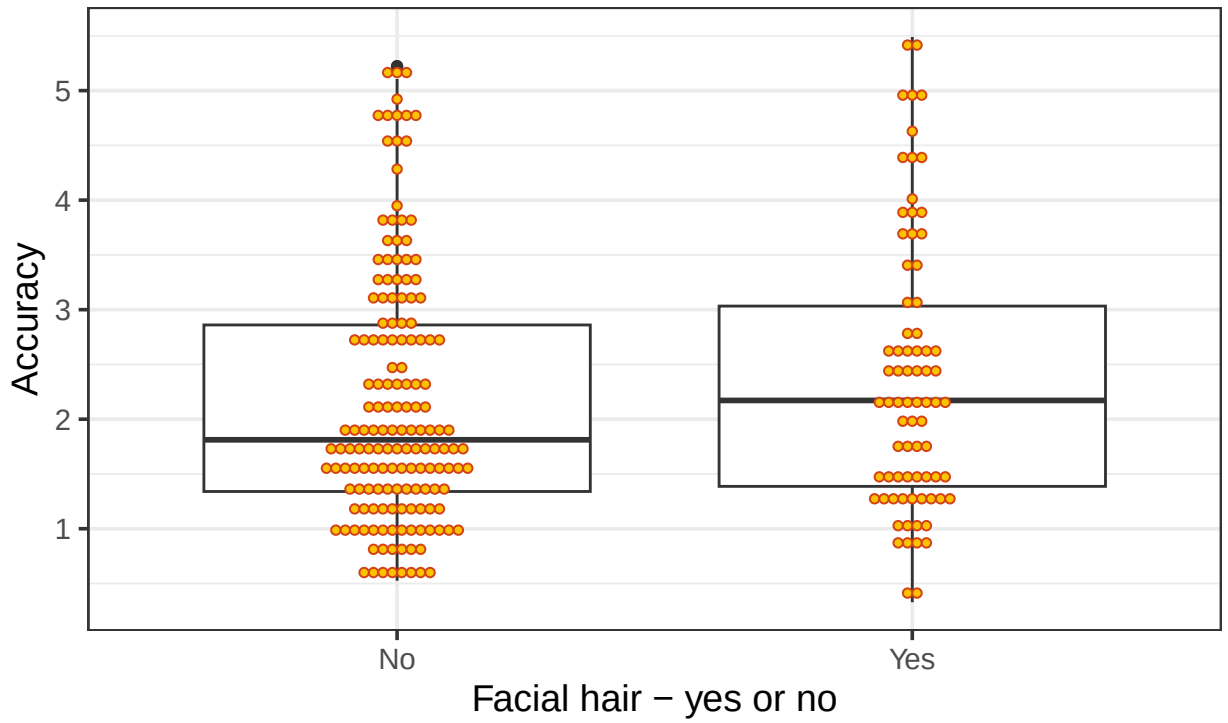
Accuracy across eye color

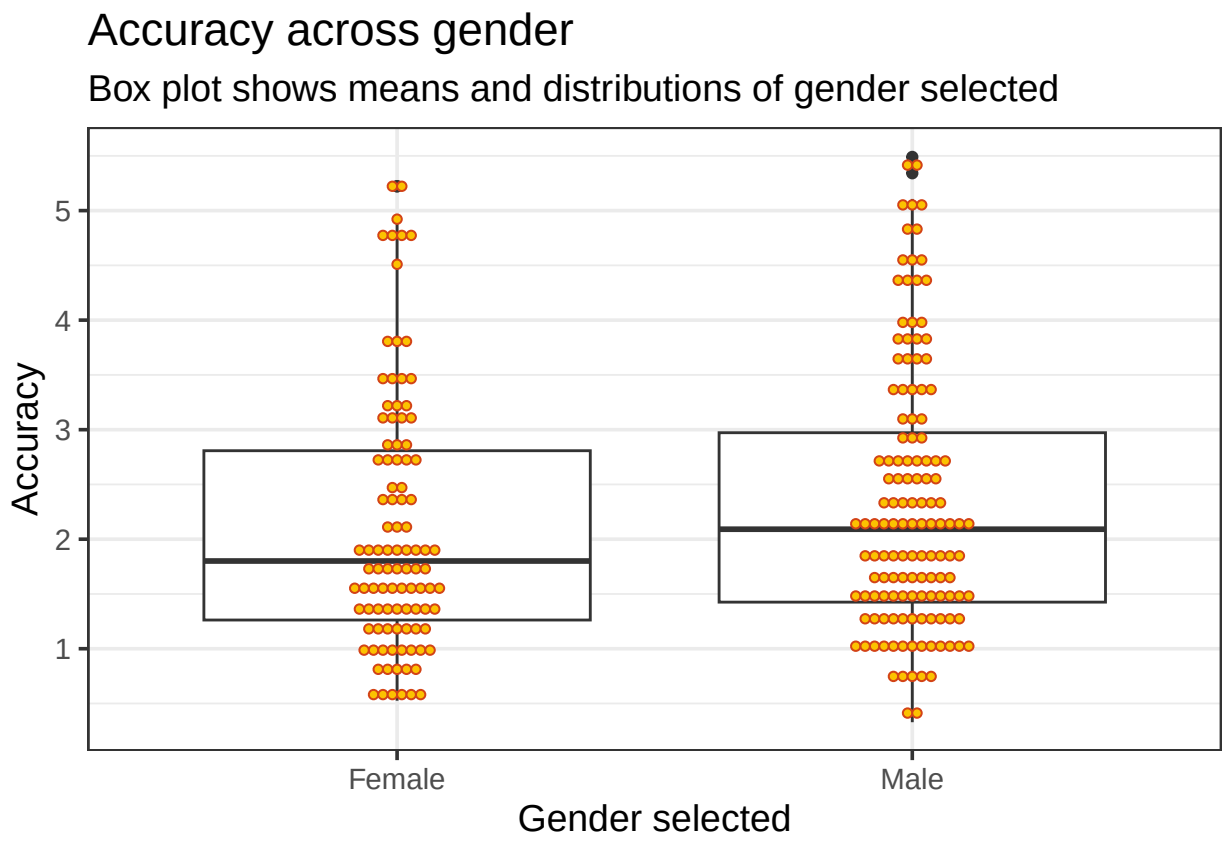
Box plot shows means and distributions of each eye color selected

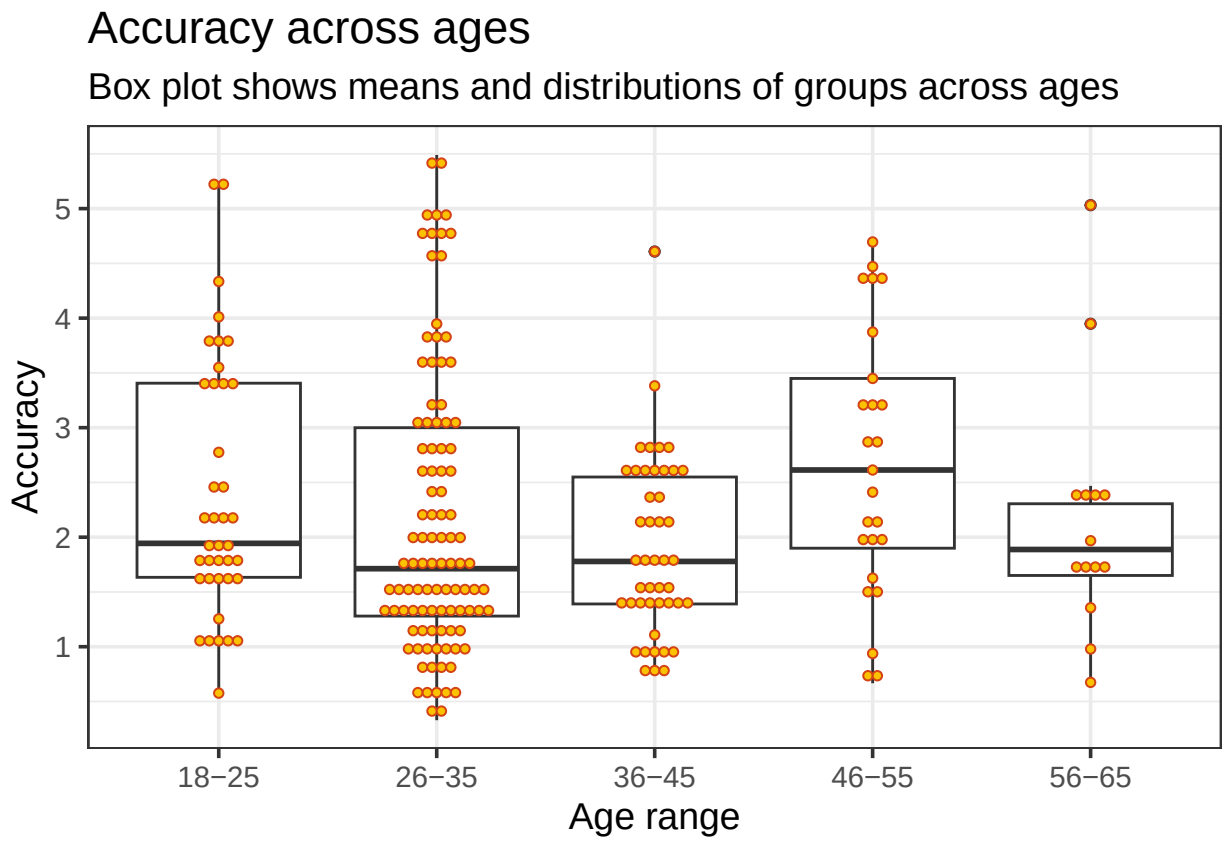


Accuracy across facial hair

Box plot shows means and distributions of groups with or without facia

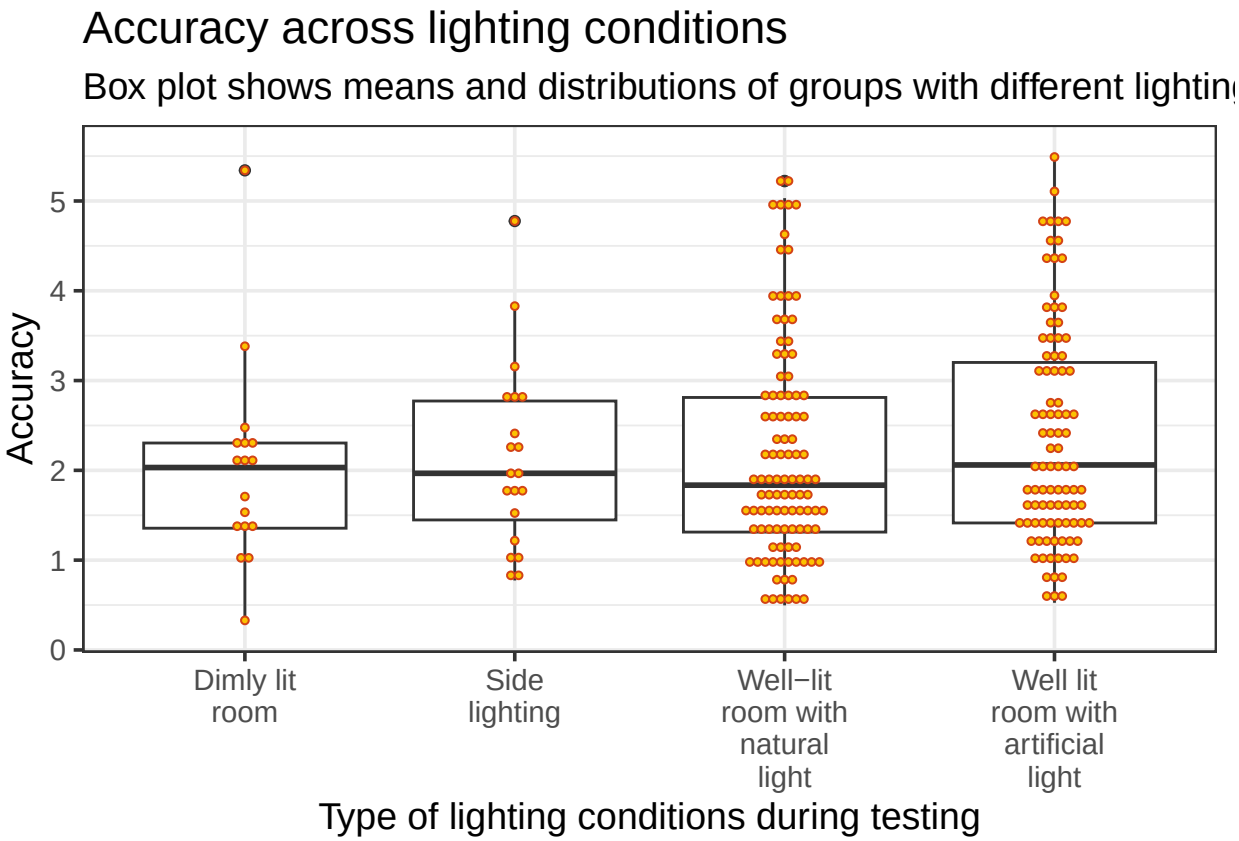



Accuracy across gender
Box plot shows means and distributions of gender selected
Accuracy
5
4
3
2
1
Female
Male
Gender selected



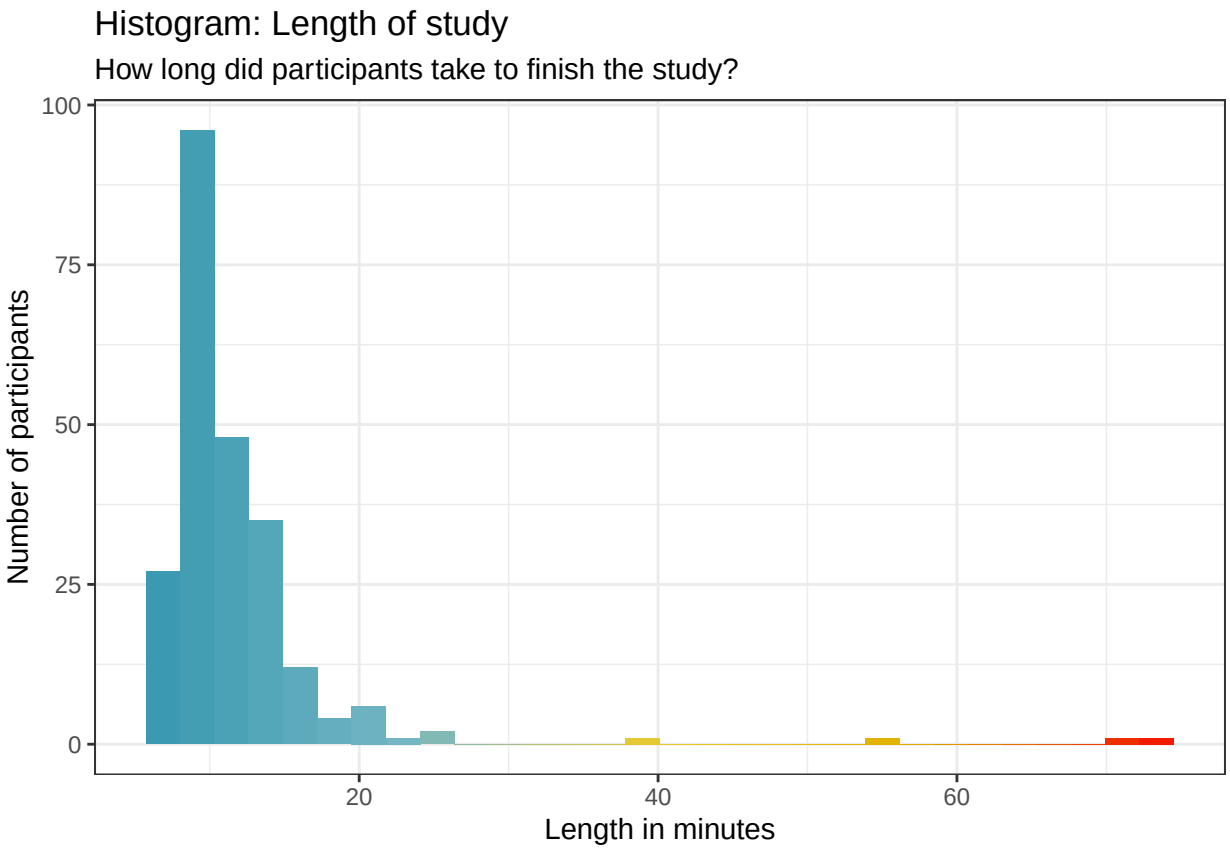
Question 3 : How much of an impact does lighting have on accuracy? The most important environment variable evaluated was self-reported lighting in the room. Only one participant reported doing the study outdoors and was removed from the analysis.

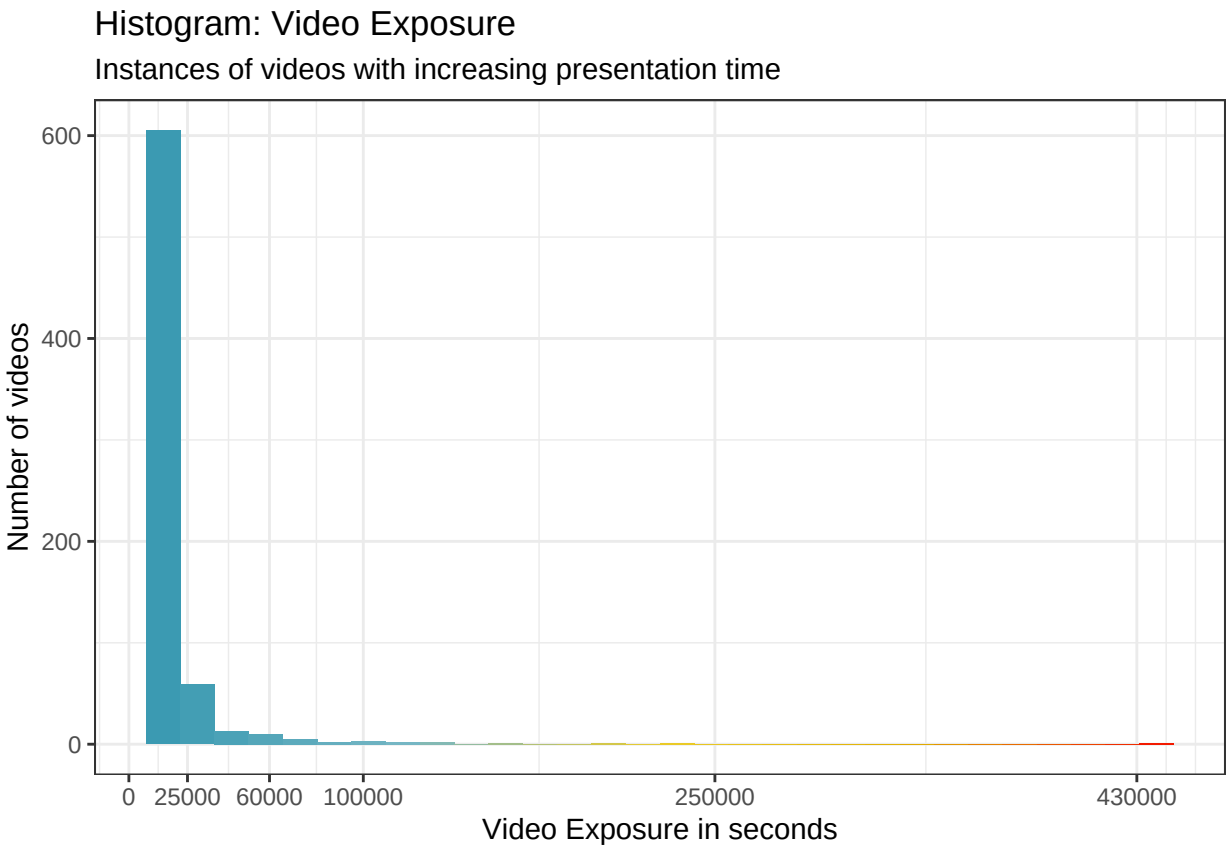
As the box plots below indicate, although there was a large variance on accuracy within self-reported lighting condition, there were no significant differences ($H(3) = 1.55$, $P = .67$) on accuracy between participants seated in different lighting conditions.

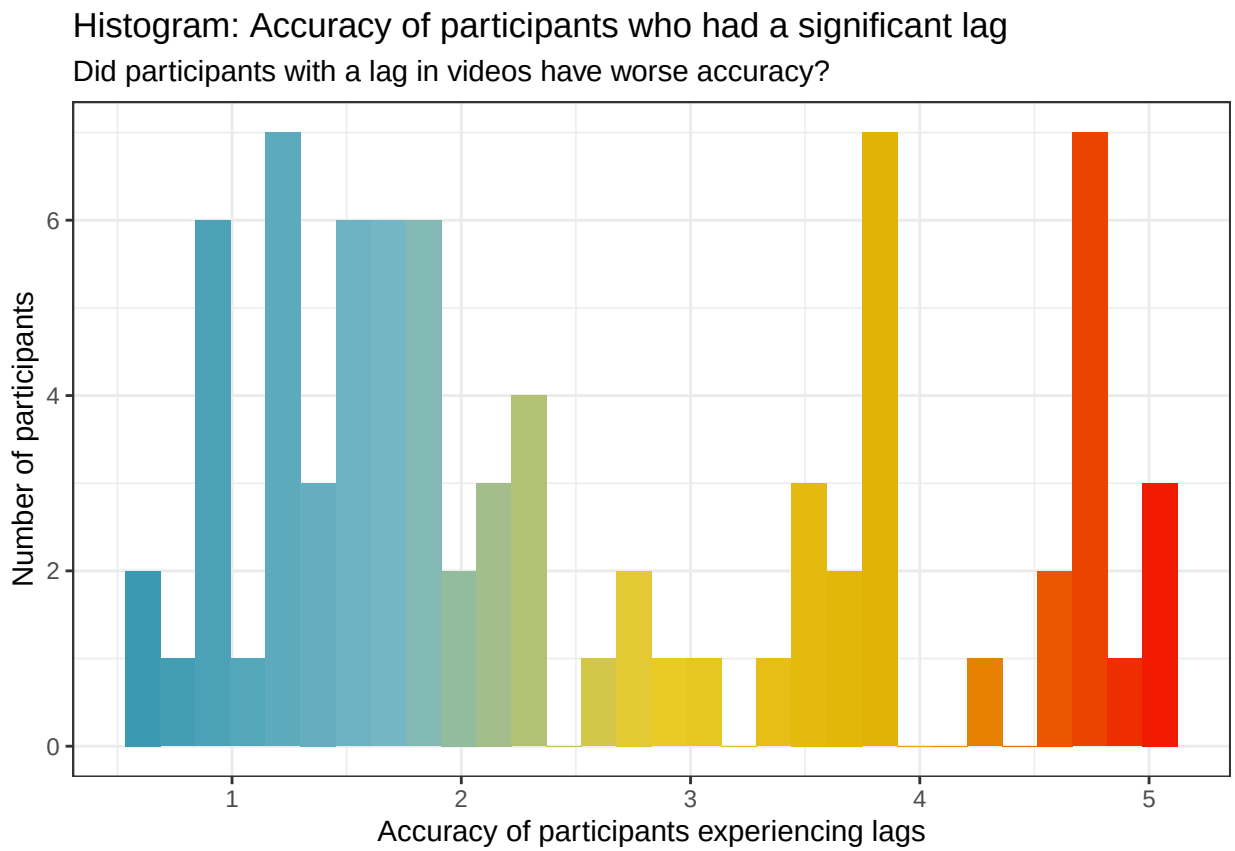


Question 4 : How does accuracy change with time? We evaluated how long each participant took, on average, to finish the study. While most participants finished the study within the stipulated 10 - 15 minutes (owing to the time needed to read instructions, complete surveys etc), there were also participants with very long study durations. This is shown in the histogram on the length of the study.

To look deeper into issues caused specifically by internet issues, we looked at those participants who had video exposures longer than 25 seconds. Since the videos were 15, 19 and 19 seconds respectively, these are participants with greater than 5 seconds of lag and are illustrated in the histogram on video exposure below. While there were definitely participants (15-20% of all videos shown) facing lag issues, the last histogram again shows the lack of correlation between study lags on videos, that resulted in less than ideal testing conditions, and the accuracy scores ($r < 0.2$)


Histogram: Length of study
How long did participants take to finish the study?
Number of participants
100
75
50
25
0
20
40
60
Length in minutes


Histogram: Video Exposure
Instances of videos with increasing presentation time
Number of videos
600
400
200
0
0
25000
60000
100000
250000
430000
Video Exposure in seconds



Question 5 : How does the accuracy translate to using AOIs? To classify whether the gaze was on target or not, a circular AOI with a diameter of 363 pixels was drawn around the target. This equals an area covering 5% of the entire stimulus. To set this in relation, a circle of this size would cover 5-6 degrees of visual angle if seen on a 14 inches laptop for participants seated around 60cms away. While the degrees in visual angle changes based on screen size and distance to the screen, the AOI would always cover 5% of the screen, as measured from the center of the AOI.

The image below shows the percentage of participants who had fixations classified within this AOI. Block 1 was presented earlier in the study, block 2 towards the end of the study.

On average people had 2.27 fixations in block 1 and 2.29 fixations in block 2.

The center of the screen seems to have the highest number participants with fixations whereas the number of participants dwindles a bit towards the bottom edges of the screen. This could be because a webcam is placed on the top of a monitor and data may be lost when people are looking down with their eyelids obstructing the webcam trying to capture the iris. From the start to the end of the study, there was a drop in how many participants could have a fixation in different locations on the screen when the emojis were shown. This could be due to people having moved during the study duration. Since the study was conducted in a real life situation, without the control of a lab or experimenter, there is no way to ascertain if this was really the case, but experimenters using the webET 3.0 with remote studies can expect similar trends.

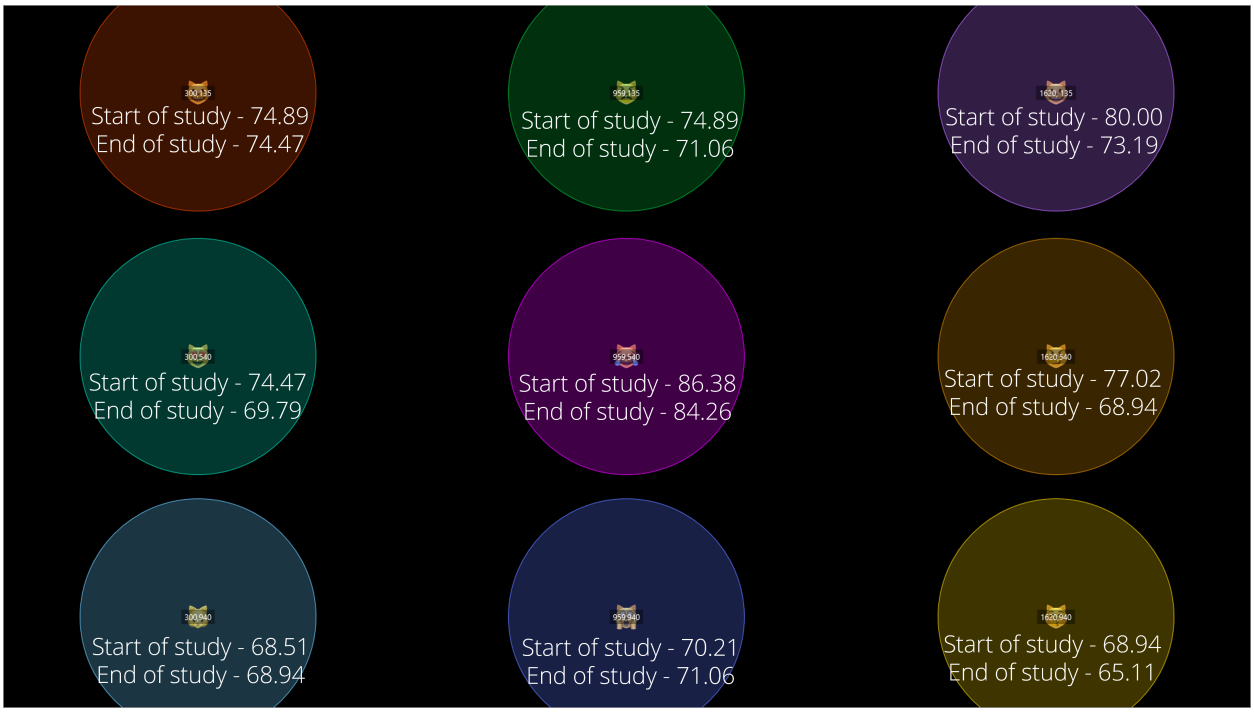
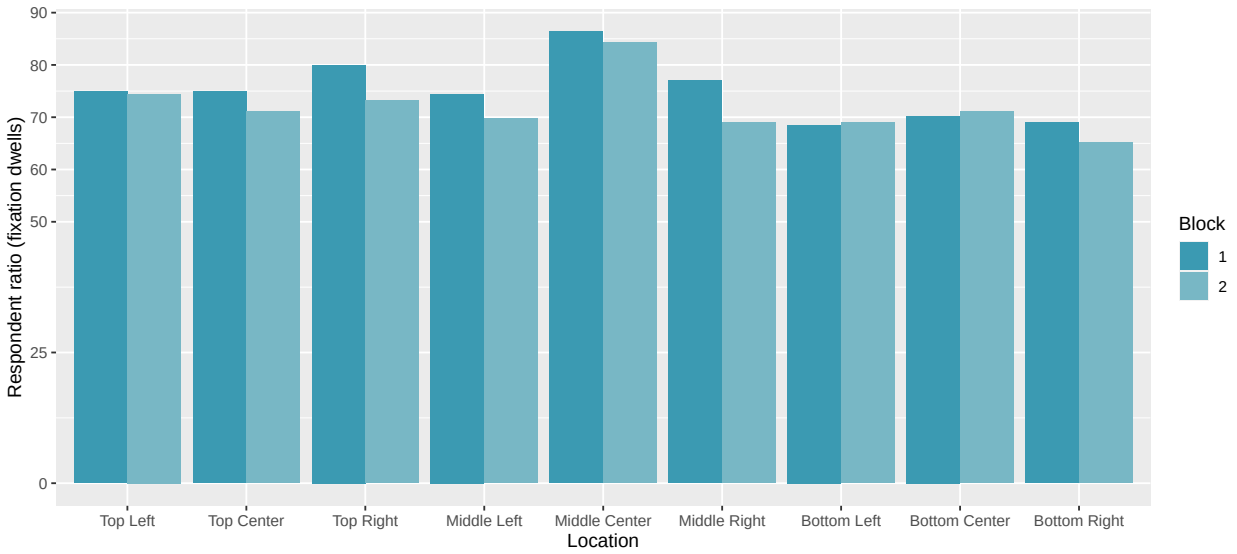


Figure 1: Percentage of participants with fixations in AOI across both blocks



4. Conclusions

1. **What is the accuracy distribution of a dataset collected with WebET 3.0?**
   If participants are prompted to have a head and eye check at the start of the study, and use the recommended 13 pre- and post- calibration slides, with inter-stimuli slides periodically through a 10-15 minute study, over 90% participants have less than 5 degrees of accuracy and over 70% have accuracy below 3dva, with the median accuracy being 2.08.
2. **Do individual and demographic variables affect accuracy?**
   While individuals wearing glasses can influence WebET even on a group level, variability from other demographic variables like ethnicities, eye-color, facial hair, age and gender can be compensated for in larger datasets leading to no significant differences on accuracy.
3. **How much of an impact does lighting have on accuracy?**
   Different indoor lighting conditions were not found to have a significant impact on accuracy. However, from the data we have about self-reported room conditions, it is difficult to know how exactly faces were illuminated and illumination differences on the face can still have an impact on eye tracking.
4. **How does accuracy change with time?**
   While longer studies and internet problems can cause issues with participant compliance and a sub-optimal user experience, the time taken to complete the study by itself does not seem to be related to accuracy calculated via calibration slides. However, fewer and less accurate data may be collected as time passes (refer to point 5 below).
5. **How does the accuracy translate to using AOIs?**
   The number of participants for whom fixations can be detected is highest in the center of the screen and reduces towards the lower corners of the screen. Over the course of the study, there seems to be a drop in the number of participants for whom gaze can be identified (and as a result fixation can be classified) on the target locations. This could be because compliance reduces with a bad user experience and researchers are still advised to keep studies as short as possible.