RealEye sp. z o.o.

www.realeye.io

May 2025

v. 3.1.1

RealEye Webcam Eye-Tracking

for Desktops, Laptops and Smartphones

Technology White Paper

Authors:

Martyna Pietrzak – MA (Cognitive Sciences), martyna@realeye.io

Andrew Duchowski – PhD (Computer Science), Professor, Clemson University, duchowski@clemson.edu

Izabela Krejtz – PhD, Associate Professor, SWPS Univeristy, ikrejtz@swps.edu.pl

Krzysztof Krejtz – PhD, Associate Professor, SWPS Univeristy, kkrejtz@swps.edu.pl

Filip Żarnowiec - BEng (Biomedical Engineering), filip@realeye.io

Damian Sromek - BEng (Applied Informatics), damian@realeye.io

Table of contents

Glossary **5**

1. Introduction **7**

2. Background: webcam eye-tracking solutions validated in non-laboratory settings **8**

2.1. Desktops and Laptops (Computers) **8**

iMotions WebET 8

Sticky by Tobii 9

WebGazer 9

2.2. Smartphones **10**

Sticky by Tobii 10

iTracker 11

3. RealEye webcam eye-tracking technology **11**

3.1. Fixation estimation pipeline **12**

3.2. Fixation filter **13**

3.3. Virtual chinrest **13**

4. Validation methodology **14**

4.1. Evaluation metrics **14**

4.2. Study design and flow **15**

Fixation target 16

4.3. Data processing pipeline **17**

Data cleaning 18

Data preprocessing 18

5. Accuracy on Desktops and Laptops **19**

5.1. Experimental setup **19**

Calibration 20

Re-validation 21

Large Grid 21

5.2. Participants **21**

5.3. Results **22**

Accuracy and its decay over time 22

Full-Screen Accuracy – Large Grid task 25

Comparison of RealEye results on Computers with other approaches 30

6. Accuracy on Smartphones 32
6.1. Experimental setup 32
Calibration 32
Re-validation 33
Large Grid 33
6.2. Participants 34
6.3. Results 34
Accuracy and its decay over time 34
Full-Screen Accuracy – Large Grid task 37
Comparison of RealEye results on Smartphones with other approaches 42
7. Conclusions and Implications 43
7.1. Desktops and Laptops – key findings 43
Accuracy and its decay over time 43
Full-screen Accuracy 43
Survey-Based Analysis 44
7.2. Smartphones – key findings 44
Accuracy and its decay over time 44
Full-screen Accuracy 44
Survey-Based Analysis 45
7.3. Limitations and Implications for future research 45

Glossary

Fixation	A series of gaze points that are very close in terms of time and space. To determine whether a cluster of gazes is a fixation, *minimum fixation duration*, *noise reduction*, and *gaze velocity threshold* parameters need to be defined. See more: https://support.realeye.io/fixation-filter
Gaze (point)	A point of where a person is looking (a sample of the eye-tracking data); webcams usually have a 30Hz sampling rate, which means that for webcam eye-tracking, there are ~30 gaze points each second.
Gaze velocity threshold	A velocity "limit"; data below this limit is classified as *fixation*, while data that exceeds this velocity limit is a *saccade*. This filter may be set to a 0 value while having set a suitable *noise reduction* level.
Minimum fixation duration	A minimum time of gaze cluster (pause of the eye) that may be called a fixation. It typically ranges from 100 to 300 milliseconds.
Noise reduction (window)	A time window (TOI - Time of Interest) used to denoise the raw data with the moving median algorithm. Each point is replaced by a median of points from the TOI.
Saccade	A movement of a user's eyes between fixation points, during which what is seen is not processed.

1. Introduction

Eye-tracking is a valuable tool for understanding user attention, behavior, and interaction with digital interfaces. By capturing data on where, when, and how long a user's gaze is focused, eye-tracking enables a range of applications in user experience (UX) research[1], accessibility[2] , education[3], and market analysis[4]. Historically, eye-tracking required specialized hardware – typically infrared-based cameras and fixed setups[5] - which, while accurate, limited accessibility and scalability. In recent years, advancements in computer vision and machine learning have led to the development of webcam-based eye-tracking systems that broadened its availability on desktops, laptops, tablets and smartphones[6,7].

While there exists a test battery evaluating the Accuracy for remote/desktop[8] eye-trackers, it has been only partially adapted for webcam-based eye-tracking[9]. Researchers yet employ varied methodologies to validate their solutions, which complicates comparisons – particularly when some systems are assessed solely in controlled laboratory environments. This emphasizes the importance of establishing standardized methods to ensure reliability across different solutions.

Following an overview of available webcam eye-tracking solutions validated in non-laboratory settings, this white paper presents the technical specifications of RealEye (https://www.realeye.io/) and explores various Accuracy calculation methods employed in previous studies. It evaluates the reliability of RealEye under natural conditions including user-specific variables across Desktops and Laptops (Computers), and Smartphones. The results produced by each approach are then compared to facilitate more informed decision-making when selecting among available solutions.

[1] Novák, J. Š., Masner, J., Benda, P., Šimek, P., & Merunka, V. (2024). Eye tracking, usability, and user experience: A systematic review. *International Journal of Human–Computer Interaction*, *40*(17), 4484-4500.

[2] Majaranta, P. (Ed.). (2011). *Gaze interaction and applications of eye tracking: Advances in assistive technologies: Advances in assistive technologies*. iGi Global.

[3] Ke, F., Liu, R., Sokolikj, Z., Dahlstrom-Hakki, I., & Israel, M. (2024). Using eye-tracking in education: review of empirical research and technology. *Educational technology research and development*, 1-36.

[4] Mansor, A. A. B., & Isa, S. M. (2018). The impact of eye tracking on neuromarketing for genuine value-added applications. *Global Business and Management Research: An International Journal*, *10*(1), 1-11.

[5] Płużyczka, M. (2018). The first hundred years: A history of eye tracking as a research method. *Applied Linguistics Papers*, (25/4), 101-116.

[6] Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). Webgazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 3839-3845).

[7] Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2176-2184).

[8] Ehinger, B. V., Groß, K., Ibs, I., & König, P. (2019). A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. *PeerJ*, 7, e7086.

[9] Kaduk, T., Goeke, C., Finger, H., & König, P. (2023). Webcam eye tracking close to laboratory standards: Comparing a new webcam-based system and the EyeLink 1000. *Behavior Research Methods*, 1-21.

2. Background: webcam eye-tracking solutions validated in non-laboratory settings

In this chapter, several available webcam eye-tracking solutions validated in natural/non-laboratory settings are reviewed, covering both Computers and Smartphones. These systems vary in technical specifications, calibration procedures, and reported accuracy metrics. While brief details on each solution's methodology and accuracy are included, the lack of a systematic validation methodology limits the ability to directly compare their reliability and effectiveness.

2.1. Desktops and Laptops (Computers)

Research shows that webcam-based eye-trackers can achieve good accuracy and reliability, often matching traditional infrared systems in specific situations, particularly in laboratory environments[9,10,11]. However, there are only several webcam eye-tracking solutions whose accuracy is validated in natural settings, each employing a slightly different methodology.

iMotions WebET[12]

iMotions employed a 13-point calibration at the beginning and end of the study, with additional inter-stimuli calibration points throughout. Cat gifs and emojis were used as validation targets, presented at 9 randomized screen locations. However, the participants' task was not clearly described, the stimulus coordinates were not provided, and the duration of the stimulus display was not specified.

The reported accuracy was approximately 3 to 5 degrees of visual angle (DVA), with 92% of participants achieving an accuracy of less than 5.5 DVA, and the median accuracy score was 2.08 DVA. This measurement was calculated as the mean Euclidean distance of all fixations. It is important to note, however, that this is only an estimate, not a precise calculation. As the authors acknowledge, "While we asked participants for their screen size, there is no way to track their distance from the screen. An assumption that works for everyone is likely to result in some unknown offsets."

[10] Burton, L., Albert, W., & Flynn, M. (2014, September). A comparison of the performance of webcam vs. infrared eye tracking technology. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 58, No. 1, pp. 1437-1441). Sage CA: Los Angeles, CA: SAGE Publications.

[11] Wisiecka, K., Krejtz, K., Krejtz, I., Sromek, D., Cellary, A., Lewandowska, B., & Duchowski, A. (2022, June). Comparison of webcam and remote eye tracking. In *2022 Symposium on eye tracking research and applications* (pp. 1-7).

[12] Seernani, D., Pedersen, M., & Wolf, K. (2023). *WebET 3.0 - Validation Study Report*. iMotions. https://doi.org/10.13140/RG.2.2.32959.07849

Sticky by Tobii[13]

The validation procedure for Tobii's Sticky involved a calibration process at the beginning, though the exact number of calibration points was not specified. Participants viewed 20 stimuli images in randomized order, each displayed for 3 seconds on a white background. The stimuli were images of faces occupying 10% of the screen height, distributed across the screen, but their exact coordinates were not provided. A countdown from 3 to 1 was displayed on the stimuli to direct participants' attention as they were instructed to fixate on the images. Of the total sessions, 63% passed the validation standards and were included for analysis.

The authors used the term "gaze error" to describe the discrepancy between the recorded gaze points and the intended target, which serves as a key metric to evaluate the system's accuracy. It was calculated as the average distance between the estimated gaze points during the last 2 seconds of the stimulus display, and the closest edge of the target stimulus area, and expressed as a percentage of screen size. The average "radial gaze error" was reported as 6% of the screen.

WebGazer[14]

WebGazer, "a self-calibrating client-side eye-tracking library written entirely in JavaScript," reported a mean error of 104 pixels, calculated as the Euclidean distance between a click location and the corresponding predicted gaze location. However, Semmelmann and Weigelt (2018)[15] employed a more common approach to accuracy evaluation and defined accuracy (or offset) as the Euclidean distance between the target position and the estimated gaze point. Their methodology began with a calibration phase involving 26 red-colored dots (50 × 50 px, #DC4A47) presented sequentially at 13 predetermined positions on the screen. Validation consisted of fixating on five randomly ordered points (at 20% positions from the top, bottom, left, and right edges, as well as the center), with fixations required to be within 200 px of the target to pass. The grid-based fixation task consisted of trials beginning with a 1,500 ms black fixation cross, followed by a 500 ms blank screen, and a blue-colored dot (50 × 50 px, RGB: 221, 73, 75) displayed for 2,000 ms. Accuracy was measured only for fixations that occurred 1,000 ms after target onset. The study reported an accuracy of 18% of screen size, equivalent to 207 pixels.

---

[13] Tobii. (2024). *White Paper: Sticky by Tobii webcam eye tracking accuracy*. Tobii. https://connect.tobii.com/s/article/White-Paper-Webcam-eye-tracking-accuracy. Accessed November 7, 2024.

[14] Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). Webgazer: Scalable webcam eye tracking using user interactions. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3839–3845. AAAI Press.

[15] Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*, 451-465.

2.2. Smartphones

The eye-tracking tests on smartphones are gaining popularity, as smartphones have become integral to daily life, with people spending increasing amounts of time using them[16]. The approaches using infrared eye-tracking on smartphone devices (involving either a fixed and stable mobile device, typically mounted onto a table, or a wearable setup, such as eye-tracking glasses) do not reflect real-world usage. However, just as with Desktops and Laptops, several webcam-based eye-tracking solutions for Smartphones have had their accuracy validated in natural settings. Each of these methods employs a slightly different methodology, with results varying depending on the device orientation.

Sticky by Tobii[17]

The validation procedure for Tobii's Sticky for Smartphones was similar to the one used for Desktops & Laptops. It involved a calibration process at the beginning, though the exact number of calibration points was not specified. Participants viewed 8 stimuli images in randomized order, each displayed for 5 seconds on a white background. The stimuli were images of faces occupying 15% of the screen width (when in vertical orientation), distributed across the screen, but their exact coordinates were not provided. A countdown from 5 to 1 was displayed on the stimuli to direct participants' attention as they were instructed to fixate on the images. The percentage of participants who met the validation standards and were included in the analysis was not provided, except for the information that only participants with high-quality data were considered.

The authors used the term "gaze error" to describe the discrepancy between the recorded gaze points and the intended target, which serves as a key metric to evaluate the system's accuracy. It was calculated as the average distance between the estimated gaze points during the last 4 seconds of the stimulus display, and the closest edge of the target stimulus area, and expressed as a percentage of screen size. The average radial gaze error was reported as 68.92 pixels for vertical, and 57.33 pixels for portrait orientation.

---

[16] Statista. (2023). *Mobile network subscriptions worldwide 2028*. Statista. Retrieved November 21, 2024, from https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

[17] Tobii. (2024). *White Paper: Sticky by Tobii mobile eye tracking accuracy*. Tobii. https://connect.tobii.com/s/article/White-Paper-Mobile-eye-tracking-accuracy. Accessed November 21, 2024.

iTracker[18]

iTracker is a deep convolutional neural network developed using the GazeCapture dataset, which contains over 2.4 million frames from 1,474 participants, with diverse head poses, lighting conditions, and device orientations, exclusively collected from iPhone users through a specially designed application. The dataset was used for end-to-end training, validation, and testing, where the system learned directly from input data (frames from recordings), making it a simulation of real-world conditions rather than an actual field test. During data collection, participants were instructed to fixate on dots with 60 random, and 13 fixed locations on the screen. The data from 13 fixed red dots locations for each subject was used to simulate the process of calibration.

The accuracy of iTracker was assessed by comparing predicted gaze positions to known target locations on the screen, with accuracy measured as the Euclidean distance in centimeters. Fine-tuned for each device and orientation, the model achieved a "prediction error" (their measure of accuracy) of approximately 1.34 cm with "calibration" and 1.56 cm without "calibration".

3. RealEye webcam eye-tracking technology

RealEye is an online platform designed for conducting webcam-based eye-tracking research. It allows researchers to gather insights into participants' visual behavior using standard desktops, laptops, or mobile devices. As a comprehensive solution, RealEye enables users to create personalized studies, collect eye-tracking and facial coding data, and analyze the results in an intuitive, user-friendly interface. The platform runs in a web browser, eliminating the need for external hardware and software installations, therefore providing a cost-effective approach to studying visual attention patterns in images, videos, websites, and more, and enabling broader exploration of human-computer interactions across diverse environments. With support for multiple languages, RealEye is accessible for international research studies.

To conduct a study using RealEye, a webcam and an internet connection are required. Built-in cameras on most laptops and smartphones are generally sufficient. However, for desktop computers, it is recommended to use a webcam with a resolution of at least 1080 pixels at 30 frames per second, ensuring a minimum frame rate of 20 frames per second. RealEye platform supports eye-tracking tests

[18] Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2176-2184).

at a sampling rate of 60 Hz, requiring a webcam capable of recording at 60 FPS to achieve this. For smartphones, studies must be conducted with the phone held vertically and with a front-facing camera. It is advisable that the device runs an up-to-date version of the operating system (Windows, macOS, or Linux for computers, and iOS or Android for smartphones), along with a modern web browser such as Google Chrome, Microsoft Edge, Mozilla Firefox, Opera, or Safari. Keeping both the operating system and browser updated ensures compatibility with the latest features and security protocols, which is crucial for optimal performance. For further details on the recommended setup, refer to the RealEye System Requirements[19].

To try RealEye on Computers, visit http://www.realeye.io/demo/desktop/images-and-videos or click the button below. To experience RealEye on Smartphones, visit http://www.realeye.io/demo/mobile/images-and-videos or scan the QR code below.

Try RealEye Demo
for Desktops and Laptops

Scan QR code
to **try RealEye Demo**
for Smartphones

3.1. Fixation estimation pipeline

RealEye's fixation estimation pipeline operates through two modules: *Image Analysis* and *Gaze Estimation*. The former uses a machine learning model to analyze the incoming images captured by the participant's webcam, detecting eye positions by identifying key features, such as pupils and corners of the eyes. These isolated regions serve as input data for the subsequent Gaze Estimation Module, allowing the system to establish a baseline for where the participant's gaze is directed.

Once the eye positions are identified and key areas isolated, the Gaze Estimation Module employs linear regression and other predictive algorithms to estimate the participant's gaze position with greater accuracy. During calibration, the system uses predefined points and backgrounds to fine-tune the model. Calibration ensures the system can adapt to individual differences in eye movements and visual focus, enhancing the overall accuracy of gaze prediction.

[19] https://support.realeye.io/webcam-eye-tracking-hardware-and-software-requirements/

The fixation estimation pipeline runs directly on the participant's device, ensuring all data processing occurs locally and reduces delays in obtaining gaze predictions. This localized approach benefits both performance and privacy, as it eliminates the need to transmit raw video data to external servers, keeping all sensitive information securely on the participant's computer. Only the calculated gaze predictions are stored in a secure text format like *Timestamp: 10, GazePointX: 76, GazePointY: 24*, with gaze positions estimated as percentage values of the displayed item. RealEye's solution complies with EU regulations (GDPR)[20].

3.2. Fixation filter

RealEye employs a method similar to the I-VT (Velocity-Threshold Identification) fixation filter for analyzing the gaze data[21]. In this validation study, fixation filter required a gaze to remain steady within a defined area for 100 to 3000 milliseconds to be considered a fixation, since fixations below 100 ms might be too brief for conscious visual processing and can be considered artifacts or noise in this types of studies[22]. Additionally, a 200 ms noise reduction window was applied to refine raw data, using the moving median algorithm and replacing each point with the median of nearby points within the Time of Interest. Furthermore, 150%/s gaze velocity threshold and a maximum saccade duration of 150 ms were selected.

3.3. Virtual chinrest

Although participants are encouraged to keep their heads still for optimal eye-tracking accuracy, this is not strictly necessary and can be challenging to control in real-world, non-laboratory settings. To allow for more natural behavior, RealEye developed a "virtual chinrest" – a software solution that adjusts eye-tracking calculations to accommodate subtle head movements. It provides real-time feedback by displaying green dots on the screen whenever there's a noticeable shift in the participant's head position, indicating the recommended eye positions for maintaining stable head alignment (see Figure 1). Consequently, the virtual chinrest enhances the overall user experience by increasing comfort levels while maintaining accuracy.

[20] https://support.realeye.io/gdpr-compliance/

[21] https://support.realeye.io/fixation-filter

[22] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. oup Oxford.

Figure 1. *The Virtual Chinrest initializes after a change in the participant's position compared to the calibration process. The study continues once the participant's eyes remain fixed on the green dots for ~5 seconds.*

4. Validation methodology

RealEye collects data from standard webcams in various environments, providing flexibility and easy access for users. Consequently, the accuracy validation methodology needed to address the diverse range of conditions encountered in real-world settings. Considering the accuracy validation methods proposed by other researchers in webcam eye-tracking, several modifications were necessary to ensure replicability and comparability with RealEye's solution. Key differences included the experimental setup, as the RealEye study was not conducted in rigorously controlled laboratory conditions; the calibration procedure, which is unique to each system; the lack of direct comparisons with data collected simultaneously from dedicated hardware-based eye-trackers; and the exclusion of mid-task recalibration, which could disrupt users' natural interactions. Overall, these adaptations were essential to ensure that RealEye's methodology reflects natural device usage and provides reliable results.

4.1. Evaluation metrics

Spatial accuracy and precision are key metrics used to evaluate how well the eye-tracking system estimates participants' gaze points. The "data quality can be defined as the distance [...] between the actual x_i and the measured $\hat{x}_i$ gaze positions."[23]. In the RealEye validation study, the measurements were taken in pixels – this is due to the fact that the distance of online users from the screen was not controlled, and screen size was self-reported, introducing uncertainty. And both these variables are crucial in calculating visual angle (see Figure 2).

[23] Holmqvist, K., Nyström, M., & Mulvey, F. (2012, March). Eye tracker data quality: What it is and how to measure it. In *Proceedings of the symposium on eye tracking research and applications* (pp. 45-52).

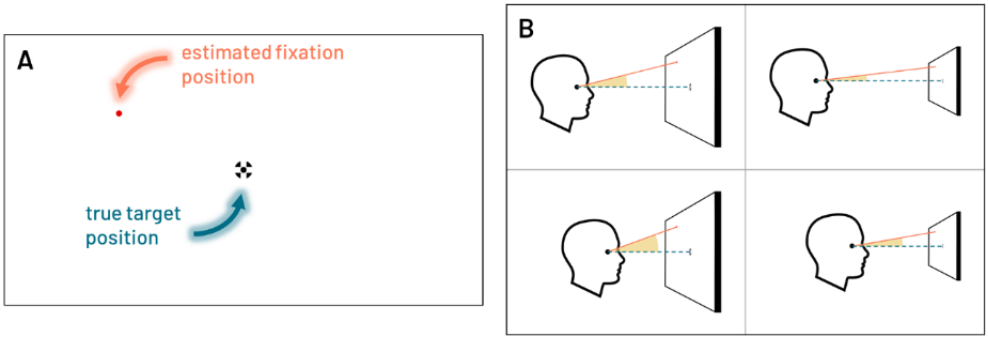

Figure 2. A. *Frontal view of the screen displaying the target and the estimated fixation position.* **B.** *Changes I visual angle (yellow) based on participant's distance from the screen and the screen size.*

To validate the RealEye webcam-based eye-tracking solution, the following metrics were used:

- **Accuracy [px]** ("offset" / "measurement error") – defined as the Euclidean distance between the center of the estimated fixation and the true target.
- **Precision [px]** – measured as the standard deviation of the selected samples.

Since data quality can vary across different areas of the screen[24], the median accuracy and precision were first calculated for each target individually. The overall metrics were then determined by averaging the medians / standard deviations across all targets. Precision, however, is considered less critical than accuracy in this case, as only a single fixation/measure was taken for each participant, and precision is typically more useful when examining within-participant variance (which was not done in this study).

4.2. Study design and flow

The design of the study was primarily inspired by Ehinger et al.[8], but shortened, as it was determined that this modified design would be sufficient, since previous webcam-based eye-tracking studies also focused on grid-based fixation tasks.

The test was created in the RealEye Dashboard[25] as an image-based study, and participants accessed it via a designated link. The study flow was identical for both Desktops & Laptops, and for Smartphones. It consisted of an automatic (fixate-only) Calibration, followed by 3 grid based tasks with no breaks in between: First Re-validation, Large Grid, and Second Re-validation (see Figure 3).

[24] Holmqvist, K. (2017). Common predictors of accuracy, precision and data loss in 12 eye-trackers. In *The 7th scandinavian workshop on eye tracking* (pp. 1-25).
[25] https://www.realeye.io/features/online-analysis-dashboard

Figure 3. *Visual representation of the study flow.*

The Re-validation task was created for two main purposes: (1) to measure accuracy immediately after Calibration, and (2) to assess whether accuracy declines over time. The task was therefore performed twice: first, following the Calibration (First Re-Validation), and again after completing the Large Grid task (Second Re-Validation). The Large Grid, while sharing the same core objective as the Re-validation task, allowed for accuracy measurement across additional, wider range of areas of the screen.

Following the three tasks, participants were presented with a Survey, during which the eye-tracking data was not recorded. Respondents selected their eye color and answered whether they used any vision aids during the experiment. Additionally, participants were asked to provide their race/ethnicity to help understand potential demographic influences, and to select one from given descriptions of the surrounding lighting conditions. Finally, the users were thanked for their participation, and the study was completed. The entire session lasted around 5-6 minutes.

Fixation target

A visual target combining a bull's eye and a cross was selected, as it was used by Ehinger et al.[8] and was proven to provide the highest fixation stability[26] (see Figure 4). It was used in the Calibration, as well as in the subsequent tasks to maintain consistency in visual stimuli throughout the experiment. The target's size and dimensions were determined in alignment with the Web Content Accessibility Guidelines (WCAG) 2.2, which advise that "the size of the target for pointer inputs is at least 24 by 24 CSS pixels"[27]. The inner circle of the target was therefore set to a diameter of 24 CSS pixels, and the overall diameter was 72 CSS pixels. The target was always presented with a contrast ratio of at least 3:1 against adjacent colors as recommended by WCAG 2.2[28]. A 14x14 CSS pixel cross was added within each target to help users with fixating on the target center.

[26] Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision research*, *76*, 31-42.

[27] World Wide Web Consortium (W3C). (2023). *Web Content Accessibility Guidelines (WCAG) 2.2*. https://www.w3.org/TR/WCAG22/. Accessed: November 7, 2024.

[28] World Wide Web Consortium (W3C). (2023). *Understanding Success Criterion 1.4.11: Non-Text Contrast*. https://www.w3.org/WAI/WCAG22/Understanding/non-text-contrast.html. Accessed: November 7, 2024.

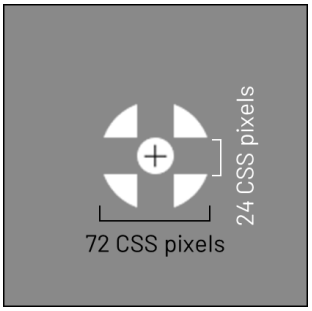

Figure 4. *A fixation target used in RealEye's system validation tasks.*

In RealEye, Calibration is performed in full-screen mode, where the fixation target is positioned based on the user's screen size while maintaining a fixed size of 72 CSS pixels. For grid-based tasks, target boards were created in Figma[29] with fixed dimensions of 1920×1080 CSS pixels for computers – the most common desktop resolution[30] – and 342×587 CSS pixels for smartphones, each with a single 72 CSS pixel target. The smartphone target board dimensions were calculated based on the average browser display area (excluding the address bar and navigation pane) on various devices.

4.3. Data processing pipeline

For both Computers and Smartphones, the data processing and analysis was the same and its flow is presented in Figure 5. The analyses were performed using Python 3.12.4[31], with the following packages and libraries versions: NumPy=1.26.4, pandas=2.2.2, SciPy=1.13.1, statsmodels=0.14.1, Matplotlib=3.9.0 and seaborn=0.13.2.

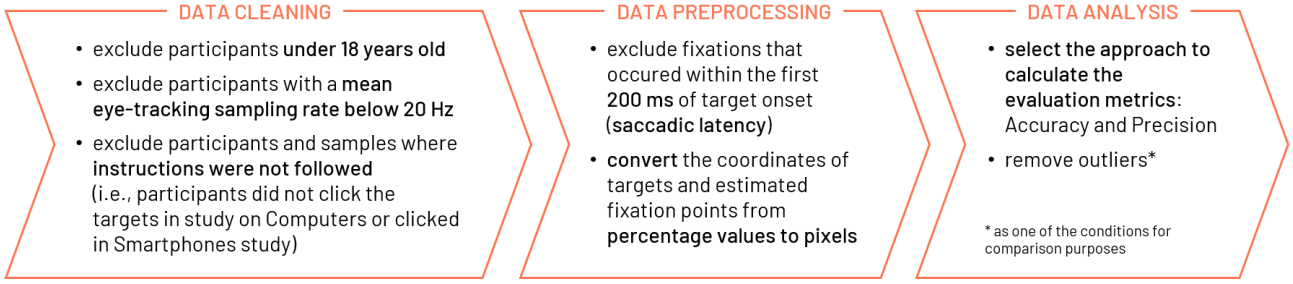

Figure 5. *A schema of data processing pipeline.*

[29] https://www.figma.com/

[30] StatCounter. (2024). *Desktop screen resolution stats worldwide*. https://gs.statcounter.com/screen-resolution-stats/desktop/worldwide (Accessed: November 7, 2024)

[31] https://www.python.org/

Data cleaning

Participants were filtered based on several criteria. First, those under 18 were excluded. Next, participants with an eye-tracking (ET) sampling rate below 20 Hz were removed, as such low frequencies may not provide enough detail for accurate analysis[21]. Finally, participants and samples that didn't follow the task instructions were excluded. This included those who didn't click or fixate while clicking in the Computer study, or those who clicked in case of the Smartphone study.

Data preprocessing

Fixations occurring within 200 ms of target onset were excluded to account for saccadic latency – the delay in eye movement from one point to another[32,33]. Next, the coordinates of the estimated fixation points, initially recorded as percentages relative to the displayed target-board's dimensions, were converted into pixel values based on each participant's screen size. A simplified illustration of this process is shown in Figure 6.

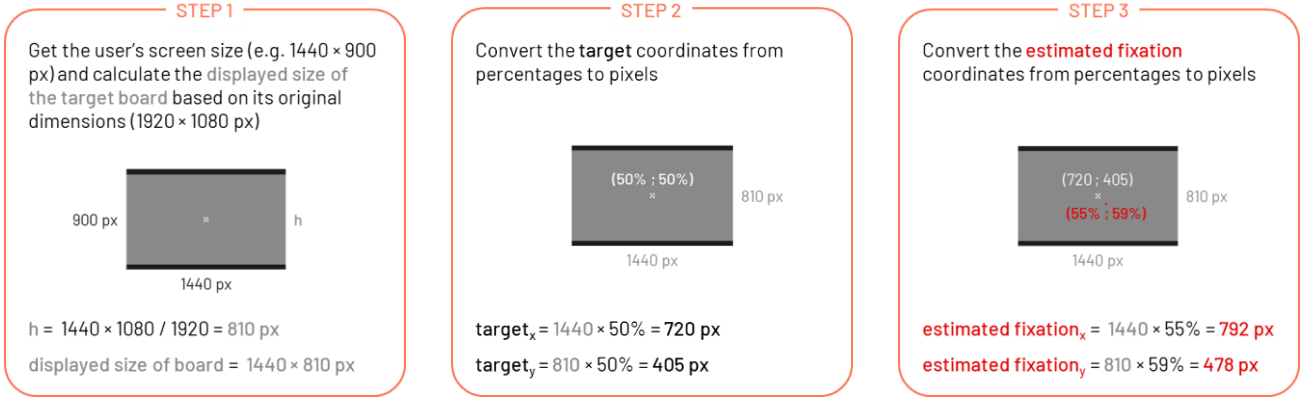

Figure 6. *A simplified overview of the process for converting data from percentage values to pixels.*

Since the method for selecting fixations for accuracy calculations is claimed not to be stringent[24], RealEye explored multiple approaches to validate its system accuracy. This choice, driven by the absence of a standardized framework, enabled an estimated comparison with methodologies used in prior webcam-based eye-tracking studies.

[32] Nyström, M., Andersson, R., Holmqvist, K., & Van De Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior research methods*, 45, 272-288.

[33] Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America*, *57*(8), 1030–1033. https://doi.org/10.1364/JOSA.57.001030

The accuracy calculation techniques included:

(1) mean Euclidean distance of all fixations from the target center[12],
(2) Euclidean distance to the average fixation centroid[9],
(3) Euclidean distance to the last fixation[8],
(4) Euclidean distance to the longest fixation[34],
(5) Euclidean distance to the fixation closest to the target center[32].

Additionally, in one of the experimental conditions, outliers were identified and removed using the Median Absolute Deviation (MAD) method[35], as the obtained data was positively skewed and heavy-tailed. This provided insights into the impact of outlier exclusion on the analysis, enhancing the reliability of accuracy measurements.

5. Accuracy on Desktops and Laptops

The way users engage with desktops and laptops (Computers) introduces unique challenges for eye-tracking systems. On desktops, users may have larger screens and more stable setups, whereas Laptops tend to be more mobile, with users frequently adjusting the angle of the screen or working in less-than-ideal postures. This variability in interaction requires RealEye's system to be adaptable, ensuring that accurate gaze data is collected regardless of how or where users position their devices. Additionally, this means the system must compensate for factors like lighting conditions, device orientation, and user movement to maintain precision.

5.1. Experimental setup

The experimental setup was based on Ehinger's test battery framework[8], using the same fixation target design and grid-based tasks for consistent evaluation of accuracy. The study was conducted with RealEye in version 15.58.13.

[34] Dalrymple, K. A., Manner, M. D., Harmelink, K. A., Teska, E. P., & Elison, J. T. (2018). An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Frontiers in psychology*, *9*, 803.
[35] Alhilo, T., & Al-Sakaa, A. (2023, September). Handling Noisy Data in Eye-Tracking Research: Methods and Best Practices. In *2023 International Workshop on Biomedical Applications, Technologies and Sensors (BATS)* (pp. 39-44). IEEE.

Calibration

Step 1 of the Calibration procedure on Computers involves 39 fixation targets appearing consecutively within a 13-point fixation grid (see Figure 7), which offers an optimal balance between calibration complexity and accuracy[36]. At the centers of the targets, the numbers from 39 to 1 in a 12px font are displayed in descending order. Each target pulsates to attract and hold participants' attention, starting at its original size, decreasing to 70%, then returning to its original size. The sequence begins with 13 white targets against a dark grey background (#393939), followed by 13 white targets against a medium grey background (#898989), and finally 13 black targets against a light grey background (#CECECE). This design aims to minimize afterimages and reduce the influence of monitor light intensity during eye-tracking.

The Calibration process is then followed by Step 2 – a system Validation, during which participants are to fixate on 3 targets one after another until they began to swirl and eventually explode. The targets are simultaneously presented on a medium-grey background. To pass the Validation, the gazes had to be no greater than 300 pixels away from the target center. In newer versions of the RealEye system, this threshold was reduced to 150 pixels to ensure the estimated gaze points were even more accurate.

Only when both steps – automatic Calibration and system Validation – are successfully completed is the calibration considered correct, and participants are allowed to continue to the tasks. Figure 7 shows a simplified flow of this procedure. If a position remained unverified for 5 seconds, the participant receives a notification that the Calibration is unsuccessful, along with a Try Again button to repeat the whole procedure. The whole process lasted around 45 seconds. 7/115 participants encountered an issue of some sort during Calibration in this validation study and required a re-Calibration, which was eventually successful for all of them.

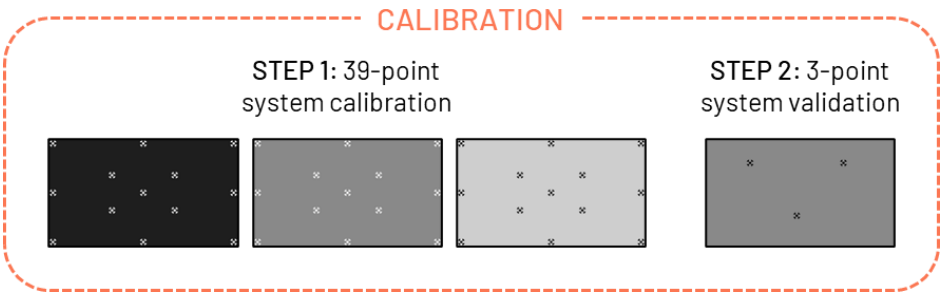

Figure 7. *Simplified visual representation of RealEye's Calibration procedure flow.*

[36] Pietrzak, M. (2024, September 24). *Optimizing RealEye calibration: Point count and accuracy on computers*. RealEye. https://www.realeye.io/blog/post/optimizing-realeye-calibration-point-count-and-accuracy-on-computers

Re-validation

During the Re validation procedure, 13 white targets appeared in a random order against a mid-grey background, starting from the center. Participants' task was to click the center of each target while fixating on it, within a 10-second period. Figure 8.A. shows a grid used in this process.

What is important, the Re-Validation procedure was not used for re-calibration purposes, but solely to provide information regarding the accuracy of system calibration. The „Re-" prefix was added to distinguish it from the system validation being part of the Calibration procedure.

Large Grid

49 white targets were positioned at equidistant intervals within a 7×7 grid across the screen and presented against a mid gray background in a random order, starting from the center. Just like in the Re-validation tasks, participants were to click on the center of each target while fixating on it, within a 10-second period. Figure 8.B. shows the visual representation of the grid used in this process.

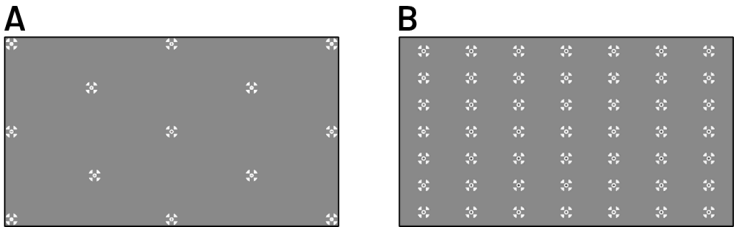

Figure 8. A. *13-point grid used in Re-validation task.* **B.** *49-point grid used in Large Grid task.*

5.2. Participants

A total of 115 participants took part in the accuracy validation study for Desktops and Laptops (Computers), with data collected in May and June 2024. 10 sessions were conducted at RealEye's headquarters in Warsaw, Poland, under semi-controlled settings, in which a desktop computer with a Windows 11 OS and 24" monitor was used. A Logitech C922 Pro Stream Webcam was placed on top of the screen, with a sampling frequency set to 30 Hz, to ensure compatibility with most popular webcams available[37].

[37] Wikipedia contributors. (2025, February 4). Webcam. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:31, February 5, 2025, from https://en.wikipedia.org/w/index.php?title=Webcam&oldid=1273962626

1 participant was excluded for being under 18 years old, 5 participants were excluded due to mean eye-tracking sampling rate below 20 Hz, and 6 participants were excluded for not following the instructions given at the beginning of the experiment. In total, 12 participants were excluded, representing 10.43% of the initial sample. An additional 1524 samples (17.44%) were excluded due to participants not clicking the target or not fixating while clicking.

The final dataset consisted of data for 103 participants, 82 of which were recruited from Prolific from UK and USA, and the remaining 21 joined the study after being directly contacted by someone from RealEye. All participants were English-speaking. Among these 103 participants, there were 49 males, 45 females, and 9 with unspecified gender. Ages ranged between 19 and 68 years (mean=37.42, std=11.21). 41% had bluish, 38% had brownish, and 21% had greenish eye colors. During the experiment, 39 respondents used vision aids, with 30 wearing glasses and 9 using contact lenses. The majority of participants were White (89), with smaller groups identifying as Asian (8) and African/African American/Black (6). In terms of lighting conditions, 74% of participants were in well-lit rooms with natural light, 15% in well-lit rooms with artificial light, 6% in dimly lit rooms, and 6% with side lighting.

5.3. Results

The findings for the RealEye webcam-based eye-tracking for Computers focus on two aspects: the accuracy immediately after the calibration and its decay over time, as well as “full-screen” accuracy measured across multiple areas of the screen. The accuracy and its decay over time were assessed based on the First- and Second Re-validation tasks, while the "full-screen" accuracy was evaluated using the Large Grid task, which measured gaze precision across a broader area of the display.

Accuracy and its decay over time

To measure the RealEye system’s accuracy immediately after calibration, the results from First Re-Validation task were calculated. To assess whether the accuracy declined over time, the results from the First Re-validation were compared with those from the Second Re-validation, which was performed as the final task of the experiment.

First Re-validation

Figure 9 shows the distribution of accuracy across participants for each accuracy calculation method, revealing a right-skewed pattern for each one of them.

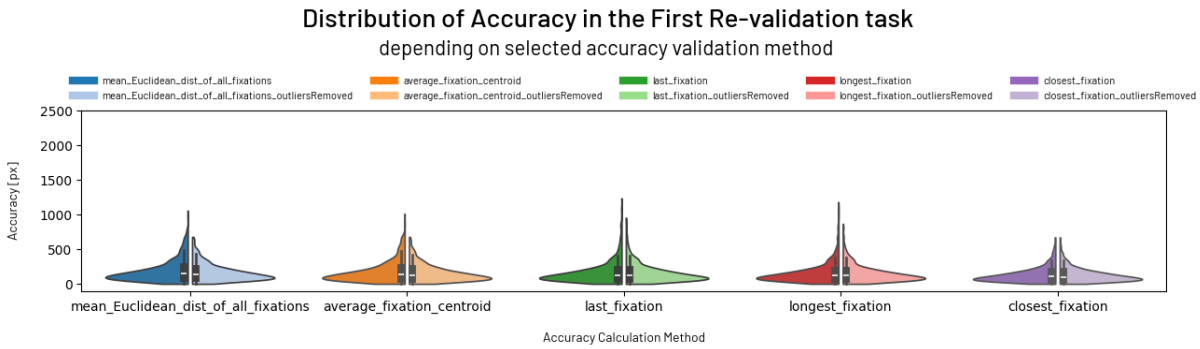

Figure 9. *Distribution of accuracy in the First Re-validation task depending on the accuracy calculation method and outliers inclusion/exclusion.*

The results presented in Table 1 show that the accuracy measured using a methodology similar to that proposed by iMotions WebET[12] yielded the worst results, with an accuracy of 149.30 pixels (increasing to 140.02 pixels when outliers were excluded). The best results were obtained using the methodology employed by Nystrom et al.[32], with calibration accuracy reaching up to 105.76 pixels (105.45 pixels with outliers removed).

Table 1. *Evaluation metrics for First Re-validation task based on different accuracy calculation methodologies.*

Method		Accuracy [px]		Precision [px]	
		outliers included	outliers excluded	outliers included	outliers excluded
(1)	mean_Euclidean_dist_of_all_fixations	149.30	140.02	147.20	116.77
(2)	average_fixation_centroid	137.61	130.02	140.91	114.70
(3)	last_fixation	122.99	122.11	143.36	123.33
(4)	longest_fixation	118.47	117.48	131.41	116.11
(5)	closest_fixation	105.76	105.45	103.94	98.98

During a viewing task, slight drifts from the intended target can occur in fixations due to factors such as eye fatigue or minor shifts in attention. By selecting only the closest fixation, the impact of this drift is minimized, allowing the focus to remain on the fixation that most accurately corresponds to the target location. Further analyses were therefore conducted using the "closest_fixation" methodology to highlight RealEye's optimal performance.

In the outlier-inclusive sample, the accuracy calculated for different areas of the screen ranged from 80 to 151 pixels, with the best results obtained in the central region (middle column) and the worst in the bottom corners. Precision was highest in the central-right part of the screen (with min = 90 pixels) and decreased towards the edges, with the lowest precision observed in the top-left corner (131 pixels).

Detailed results for accuracy and precision in different screen areas during the First Re-validation are presented in Figures 10.A and 10.B.

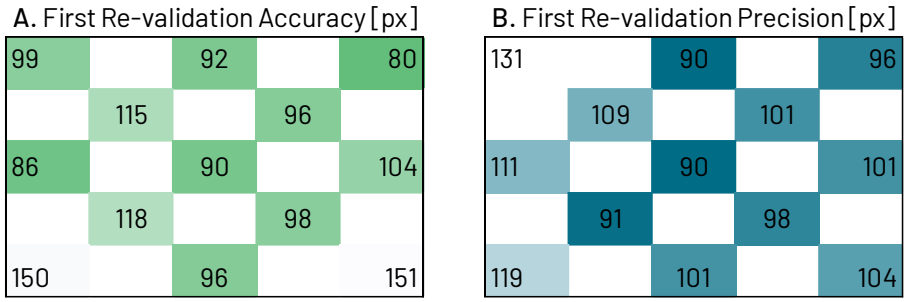

Figure 10. *"Heatmaps" of* **A.** *Accuracy and* **B.** *Precision for specific screen areas in the First Re-validation task.*

When analyzing accuracy on a per-participant-basis rather than per-item, the median accuracy was 114.13 pixels (111.45 pixels without outliers), while the precision was 79.64 pixels (52.61 pixels with outliers removed). The highest accuracy observed was 29.85 pixels, and the average accuracy for the top 30 participants was 68.80 pixels. The average accuracy for the bottom 30 participants was 235.30 pixels, which increased to 184.47 pixels after excluding participants identified as outliers using the MAD method (n = 7). Figure 11 illustrates the distribution of mean accuracies across participants, with red markers indicating the outliers. In the outlier-inclusive sample, 71.57% of participants achieved an accuracy of ≤ 150 pixels, 80.39% ≤ 200 pixels, and 91.18% ≤ 250 pixels. Without outliers, these values increased to 76.84%, 86.32%, and 97.89%, respectively.

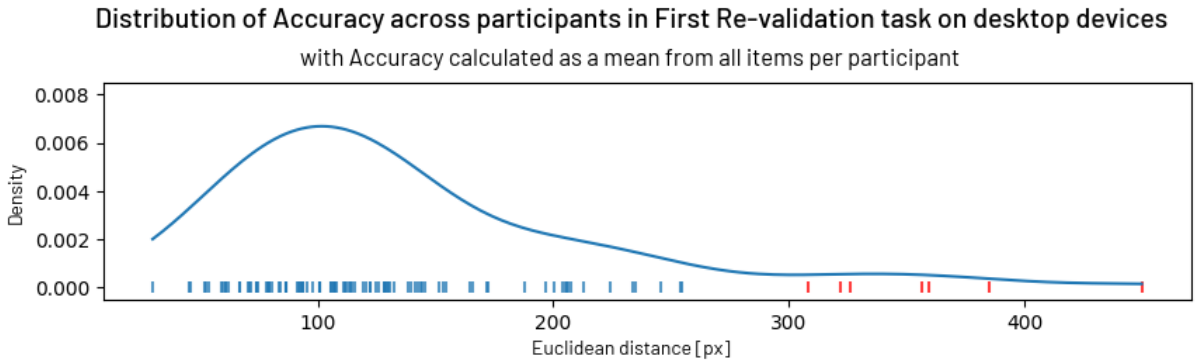

Figure 11. *Distribution of accuracy among participants in the First Re-validation task.*

Second Re-validation

To determine whether the accuracy and precision decreased over time, the metrics from the Second Re-validation were compared to those from the First Re-validation. Table 2 shows the evaluation metrics calculated for each accuracy calculation method for Second Re-validation task.

Table 2. *Evaluation metrics for Second Re-validation task based on different accuracy calculation methodologies.*

Method		Accuracy [px]		Precision [px]	
		outliers included	outliers excluded	outliers included	outliers excluded
(1)	mean_Euclidean_dist_of_all_fixations	198.34	190.92	178.58	142.32
(2)	average_fixation_centroid	188.72	179.60	174.72	140.75
(3)	last_fixation	167.36	166.18	169.27	145.63
(4)	longest_fixation	166.54	164.97	163.95	144.21
(5)	closest_fixation	146.52	145.25	146.13	128.66

For each approach, the accuracy and precision degraded over the course of the experiment. To test if the differences between the two Re-validations were statistically significant, the samples were adjusted to ensure that pairs of data from the same participants were included in both sets, and the Wilcoxon signed-rank test was applied. A statistically significant difference was revealed, with notably lower values observed in the Second Re-validation compared to the First Re-validation (p-value < 0.01). Possible explanations for this could include fatigue from repeated sessions, as the tasks were monotonous and required constant mental and visual concentration.

Full-Screen Accuracy – Large Grid task

To assess the performance of the RealEye webcam eye-tracking system for Computers on the wide range of screen areas, the Large Grid task was employed. Since the Large Grid task was not conducted immediately after Calibration, there is a possibility that accuracy results could have improved slightly if it had been, as suggested by the First Re-validation results.

Figure 12 shows the distribution of accuracy across participants for each accuracy calculation method, again revealing a right-skewed pattern for each one of them.

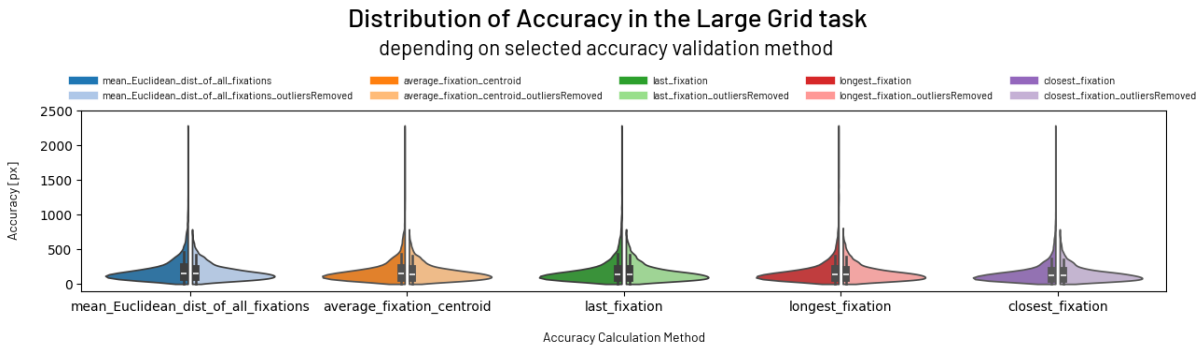

Figure 12 *Distribution of accuracy in the Large Grid task depending on the accuracy calculation method and outliers inclusion/exclusion.*

The results presented in Table 3 show that the accuracy measured using a methodology similar to that proposed by iMotions WebET[12] once again yielded the worst results, with an accuracy of 152.62 pixels (147.35 pixels when outliers were excluded). The best results were obtained using the methodology employed by Nystrom et al.[32], with calibration accuracy reaching up to 122.54 pixels (120.65 pixels when outliers were removed).

Table 3. *Evaluation metrics for First Re-validation task based on different accuracy calculation methodologies.*

Method		Accuracy [px]		Precision [px]	
		outliers included	outliers excluded	outliers included	outliers excluded
(1)	mean_Euclidean_dist_of_all_fixations	152.62	147.35	148.58	111.52
(2)	average_fixation_centroid	144.50	140.12	145.44	110.75
(3)	last_fixation	139.71	137.49	147.63	117.79
(4)	longest_fixation	135.83	134.35	146.72	113.19
(5)	closest_fixation	122.54	120.65	126.92	104.45

In the outlier-inclusive sample, the accuracy for multiple areas of the screen ranged from 98 to 165 pixels, with the best results obtained for the central-bottom part of the screen (~103 px), and the worst in the bottom-right corner (165 px). Precision was highest in the central part of the screen (with a minimum of 84 pixels), and lowest in the left-central-bottom area (264 pixels), which appears somewhat random. However, aside from this anomaly, the worst precision results were observed at the corners. Detailed results for accuracy and precision for different parts of the screen in the Large Grid task are presented in Figure 13.A and 13.B.

A. Large Grid Accuracy [px]

122	127	106	108	117	98	136
126	133	135	115	134	111	112
127	119	121	138	111	141	123
133	133	127	109	119	129	115
136	140	114	117	126	127	124
117	106	104	102	103	117	132
127	136	117	104	141	129	165

B. Large Grid Precision [px]

137	126	112	117	146	152	121
123	148	137	91	136	133	125
125	136	130	127	101	116	121
134	110	95	84	100	103	141
144	264	102	128	119	128	153
121	98	107	112	132	140	148
141	105	119	114	113	114	151

Figure 13. *"Heatmaps" of **A.** Accuracy and **B.** Precision for specific screen areas in the Large Grid task.*

When analyzing accuracy on a per-participant-basis rather than per-item, the median accuracy was 123.04 pixels (120.62 pixels without outliers), while the precision was 101.50 pixels (52.52 pixels with outliers removed). The highest accuracy observed was 59.15 pixels, and the average accuracy for the top 30 participants was 85.84 pixels. The average accuracy for the bottom 30 participants was 288.80 pixels, which increased to 195.17 pixels after excluding participants identified as outliers using the MAD method (n = 11). Figure 14 illustrates the distribution of mean accuracies across participants, with red markers indicating the outliers. In the outlier-inclusive sample, 65.05% of participants achieved an accuracy of ≤ 150 pixels, 76.70% achieved ≤ 200 pixels, and 85.44% ≤ 250 pixels. Without outliers, these values increased to 72.83%, 85.87%, and 95.65%, respectively.

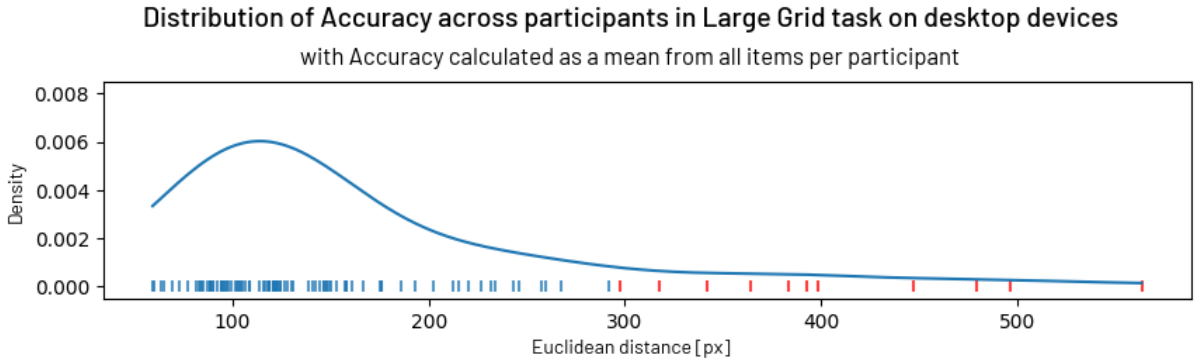

Figure 14. *Distribution of accuracy among participants in the Large Grid task.*

Survey-based analysis

To examine potential demographic and environmental factors influencing eye-tracking performance, a survey was conducted at the end of the study. It collected data on eye color, use of vision aids, race/ethnicity, and lighting conditions during the experiment. Analyzing these variables helped understand their impact on the overall accuracy and precision of the RealEye system.

Eye color

Participants self-reported their eye color by selecting one of the following options: blue, grey (Bluish), green, hazel (Greenish), amber, or brown (Brownish). Table 4 presents the evaluation metrics for each eye color group in the Large Grid task in the outlier-inclusive sample.

Table 4. *Evaluation metrics for different Eye color groups in the Large Grid task.*

Eye color	Num of participants	Accuracy [px]	Precision [px]
Bluish	42	122.53	113.34
Brownish	39	116.25	141.67
Greenish	22	136.29	110.42

To compare the results among participants, Euclidean distances were analyzed for each eye color group. The distribution of these values is shown in Figure 15. Significant differences were identified using the Kruskal-Wallis test, followed by Dunn's post-hoc test with Bonferroni-adjusted p-values. These differences were observed between the Bluish and Greenish groups, as well as between the Brownish and Greenish groups, with participants with Greenish eyes exhibiting lower accuracy in both cases (p-value < 0.05).

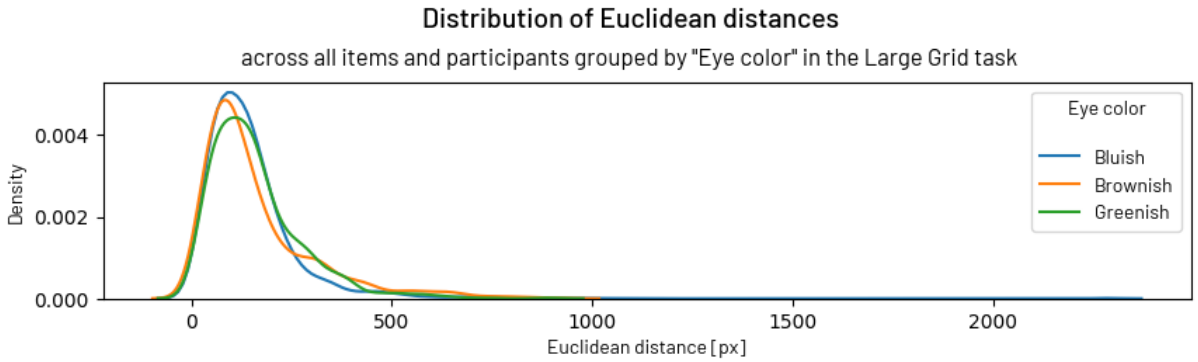

Figure 15. *Distribution of Euclidean distances calculated for specific "Eye color" groups.*

Use of vision aids

Participants reported whether they used vision aids, such as glasses or contact lenses, during the experiment. Table 5 presents the evaluation metrics for each analysis group in the Large Grid task in the outlier-inclusive sample.

Table 5. *Evaluation metrics for different Vision aids groups in the Large Grid task.*

Use of vision aids	Num of participants	Accuracy [px]	Precision [px]
No	64	118.68	120.86
Yes	39	131.41	132.23
Yes, glasses	30	138.88	143.19
Yes, contact lenses	9	113.70	78.67

To compare the results among participants, Euclidean distances were analyzed for each vision aids group. The distribution of these values is shown in Figure 16. The Kruskal-Wallis test revealed significant differences between the groups using ("Yes") and not using ("No") vision aids, particularly between participants who did not use vision aids and those who wore glasses. These findings were confirmed by Dunn's post-hoc test with Bonferroni-adjusted p-values (p-value < 0.01 in both cases). Participants who did not use any vision aids showed significantly better results than those who did, which aligns with previous research findings[24]. It is worth noting that the sample size for the "Yes, contact lenses" group was n=9, which can be considered small compared to the other groups.

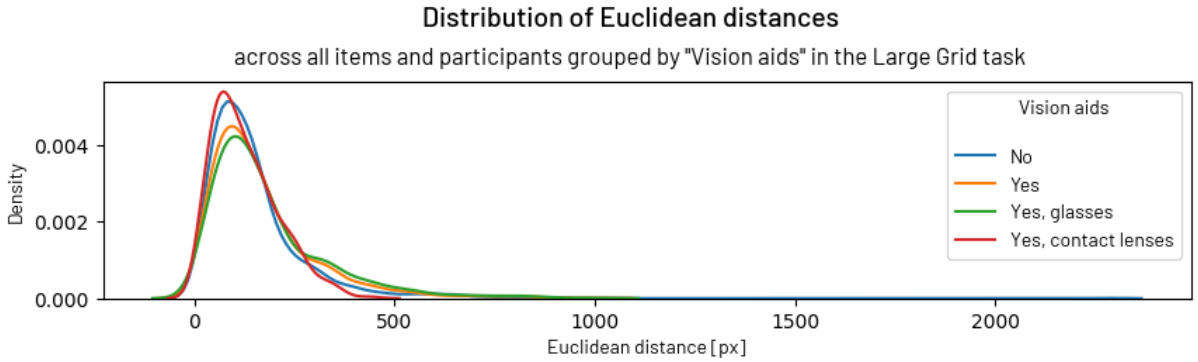

Figure 16. *Distribution of Euclidean distances calculated for specific "Vision aids" groups.*

Race/Ethnicity

Participants were asked to report their race or ethnicity by selecting one of the following: White, Asian, African / African American / Black, or Latin American / Hispanic. Table 6 presents the evaluation metrics for each analysis group in the Large Grid task in the outlier-inclusive sample.

Table 6. *Evaluation metrics for different Race / Ethnicity groups in the Large Grid task.*

Race / Ethnicity	Num of participants	Accuracy [px]	Precision [px]
White	89	117.96	165.75
Asian	8	205.23	205.49
African / African American / Black	6	165.75	118.62

Due to little amount of participants in groups of Asian and African / African American / Black, the results of ethnicity groups were not compared between each other.

Lighting conditions

Participants reported their lighting conditions by selecting one of the following options: well-lit room with natural light, well-lit room with artificial light, side lighting, or dimly lit room. Table 7 presents the evaluation metrics for each analysis group in the Large Grid task in the outlier-inclusive sample.

Table 7. *Evaluation metrics for different Lighting conditions groups in the Large Grid task.*

Use of vision aids	Num of participants	Accuracy [px]	Precision [px]
Well-lit room	91	120.44	129.82
Well–lit room with natural light	76	117.52	122.73
Well-lit room with artificial light	15	147.96	145.08
Side lighting	6	139.56	83.81
Dimly-lit room	6	135.71	96.52

Due to little amount of participants in all groups except for the well-lit room with natural light, the results were not compared between each other.

Comparison of RealEye results on Computers with other approaches

Due to the absence of a standardized framework and the variability in accuracy calculation methods, a direct comparison with previous studies is not feasible. However, to provide an approximate estimate, the accuracy data was converted from pixels to degrees of visual angle (DVA), using the Formula 1 below[38]. The DVA results were calculated based on the assumptions provided by iMotions WebET[12], including a fixed screen resolution of 1920×1080, a 14-inch diagonal screen, and an assumed viewing distance of 60 cm.

[38] Healey, C. G., & Sawant, A. P. (2012). On the limits of resolution and visual angle in visualization. *ACM Transactions on Applied Perception (TAP)*, *9*(4), 1-21.

Additionally, the average accuracy was expressed as the % of the target board width, similar to the approach used in Sticky by Tobii[13]. In their study however, the specific screen resolution(s), screen diagonal, and the assumed viewing distance were not explicitly provided. It is important to emphasize that these estimates rely on assumed variables, which could not be directly measured during the study.

Formula 1..*Mathematical equation to convert pixels to degrees of visual angle (DVA).*

$$DVA = 2 \times arctan\left(\frac{Object\ Size\ [inches]}{2 \times Distance\ [inches]}\right)$$

Where:

- $Object\ Size\ [inches] = \frac{Accuracy\ [pixels]}{PPI}$, where Pixels Per Inch $PPI = \frac{Diagonal\ Resolution\ [pixels]}{Screen\ diagonal\ [inches]}$, and $Diagonal\ Resolution\ [pixels] = \sqrt{Screen\ width^2 + Screen\ height^2}$
- $Distance\ [inches] = \frac{Distance\ from\ screen\ [centimeters]}{2.54}$

First Re-validation

In the First Re-validation task, the results indicated that 71.57% of participants achieved an accuracy of ≤ 2.31 DVA, 80.39% ≤ 3.08 DVA, and 91.18% ≤ 3.85 DVA. Without outliers, these values increased to 76.84%, 86.32%, and 97.89%, respectively. The average accuracy immediately after calibration = 105.76 px (105.45 px without outliers), is approximately 1.63 DVA (1.63 DVA), and 5.51% (5.49%) of the 1920×1080 target board width.

Large Grid

In the Large Grid task, the results indicated that 65.05% of participants achieved an accuracy of ≤ 2.31 DVA, 76.70% ≤ 3.08 DVA, and 85.44% ≤ 3.85 DVA. Without outliers, these values increased to 72.83%, 85.87%, and 95.65%, respectively. The average "full-screen" accuracy = 122.54px (120.65 px without outliers), is approximately 1.89 DVA (1.86 DVA), and 6.38% (6.28%) of the 1920×1080 target board width.

6. Accuracy on Smartphones

Using smartphones creates unique challenges for eye-tracking because users handle their phones in many different ways. Unlike desktops, where the screen stays in one position, people often change how they hold their devices, or move them closer or farther away from their faces. RealEye's system adapts to these conditions to provide reliable gaze tracking, no matter how users interact with their phones.

6.1. Experimental setup

The experimental setup was based on Ehinger's test battery framework[8], using the same fixation target design and similar grid-based tasks, but adjusted to Smartphones. The tests were performed on vertically-orientated devices only, with RealEye in version 17.4.2.

Calibration

Step 1 of the Calibration procedure on Smartphones involves 27 fixation targets appearing consecutively within a 17-point fixation grid (see Figure 17), which offers an optimal balance between calibration complexity and accuracy. At the centers of the targets, the numbers from 27 to 1 in a 12px font are displayed in descending order. Each target pulsates to attract and hold participants' attention, starting at its original size, decreasing to 70%, then returning to its original size. The sequence begins with 13 black targets against a light grey background (#CECECE), followed by 13 white targets against a dark grey background (#393939), and finally 1 white target against a middle grey background (#898989). This design aims to minimize afterimages and reduce the influence of monitor light intensity during eye-tracking.

The Calibration process is then followed by Step 2 – a system Validation, during which participants are to fixate on 4 targets one after another until they began to swirl and eventually explode. The targets are simultaneously presented in the corners on a medium grey background. To pass the Validation, the gazes had to be no greater than 150 pixels away from the target center. In newer versions of the RealEye system, this threshold was reduced to 100 pixels to ensure the estimated gaze points were even more accurate.

Only when both steps – automatic Calibration and system Validation – are successfully completed is the calibration considered correct, and participants are allowed to continue to the tasks. Figure 17 shows a simplified flow of this procedure. The whole process lasted around 30 seconds.

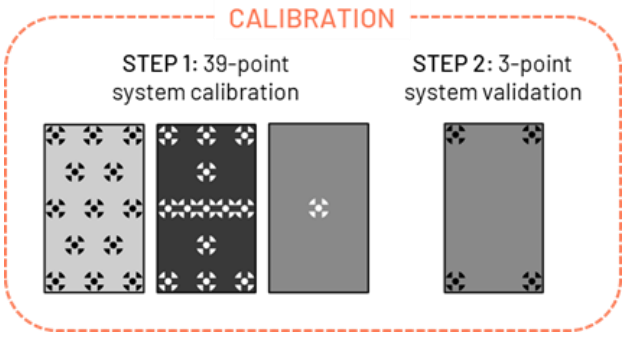

Figure 17. *Simplified visual representation of RealEye's Calibration procedure flow on Smartphones.*

Re-validation

During the Re validation procedure, 13 white targets appeared in a random order against a mid-grey background, starting from the center. Participants' task was to fixate on the center of each target until it disappeared, with the target on set for 2 seconds. Figure 18.A. shows a grid used in this process

What is important, the Re-Validation procedure was not used for re-calibration purposes, but solely to provide information regarding the accuracy of system calibration. The „Re-" prefix was added to distinguish it from the system validation being part of the Calibration procedure.

Large Grid

21 white targets were positioned at equidistant intervals within a 3×7 grid across the screen and presented against a mid gray background in a random order, starting from the center. Just like in the Re-validation tasks, participants were to fixate on the center of each, until they disappeared. Figure 18.B. shows the visual representation of the grid used in this process.

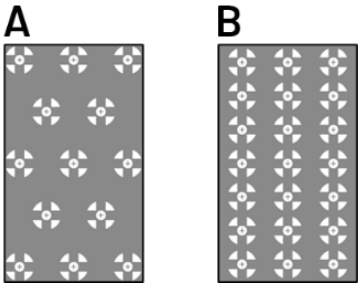

Figure 18. A. *13-point grid used in Re-validation task on Smartphones.* **B.** *21-point grid used in Large Grid task on Smartphones.*

6.2. Participants

A total of 105 participants took part in the accuracy validation study for Smartphones, with data collected in October and November 2024. Surprisingly, none of the participants had to be excluded at this point of the study, as all of them had a mean eye-tracking sampling rate above 20 Hz. However, 740 samples (14.68%) were excluded due to participants clicking the target or not fixating until the target disappeared.

The final dataset consisted of data for 105 participants, 96 of which were recruited from Prolific from UK and USA, and the remaining 9 joined the study after being directly contacted by someone from RealEye. All participants were English-speaking. Among these 105 participants, there were 40 males, 59 females, and 6 with unspecified gender. Ages ranged between 20 and 71 years (mean=39.20, std=11.59). 40% had bluish, 36% had brownish, and 24% had greenish eye colors. During the experiment, 41 respondents used vision aids, with 36 wearing glasses and 5 using contact lenses. The majority of participants were White (92), with smaller groups identifying as African/African American/Black (6), Asian (5) and Latin American / Hispanic (2). In terms of lighting conditions, 50% of participants were in well-lit rooms with natural light, 29% in well-lit rooms with artificial light, 12% in dimly lit rooms, and 10% with side lighting.

6.3. Results

The findings for the RealEye webcam-based eye-tracking for Smartphones focus on two aspects: the accuracy immediately after the calibration and its decay over time, as well as "full-screen" accuracy measured across multiple areas of the screen.

Accuracy and its decay over time

To measure the RealEye system's accuracy immediately after calibration, the results from First Re-Validation task were calculated. To assess whether the accuracy declined over time, the results from the First Re-validation were compared with those from the Second Re-validation, which was performed as the final task of the experiment.

First Re-validation

Figure 19 shows the distribution of accuracy across participants for each accuracy calculation method, revealing a right-skewed pattern for each one of them.

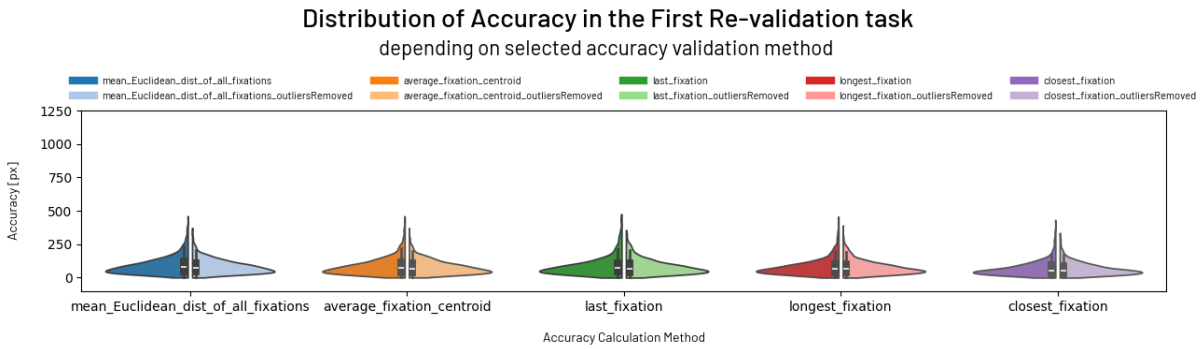

Figure 19. *Distribution of Euclidean distances in the First Re-validation task depending on the accuracy calculation method and outliers inclusion/exclusion on Smartphones.*

The results presented in Table 8 show that the accuracy measured using a methodology similar to that proposed by iMotions WebET[12] yielded the worst results, with an accuracy of 76.99 pixels (73.60 pixels when outliers were excluded). The best results were obtained using the methodology employed by Nystrom et al.[32], with calibration accuracy reaching up to 55.62 pixels (55.27 pixels when outliers were removed). Further analyses were conducted using the "closest_fixation" methodology to highlight RealEye's optimal performance.

Table 8. *Evaluation metrics for First Re-validation task based on different accuracy calculation methodologies on Smartphones.*

Method		Accuracy [px]		Precision [px]	
		outliers included	outliers excluded	outliers included	outliers excluded
(1)	mean_Euclidean_dist_of_all_fixations	76.99	73.60	61.44	49.30
(2)	average_fixation_centroid	71.48	67.74	60.84	49.84
(3)	last_fixation	69.83	68.97	67.61	54.59
(4)	longest_fixation	64.91	64.38	60.42	50.69
(5)	closest_fixation	55.62	55.27	52.50	45.31

In the outlier-inclusive sample, the accuracy calculated for different areas of the screen ranged from 45 to 75 pixels, with the best results obtained in the central-top region and the worst in the bottom corners. Precision was highest in the central-bottom part of the screen (minimum = 41 pixels) and decreased towards the edges, with the lowest precision observed in the bottom-left corner (61 pixels). Detailed results for accuracy and precision in different screen areas during the First Re-validation are presented in Figures 20.A and 10.B.

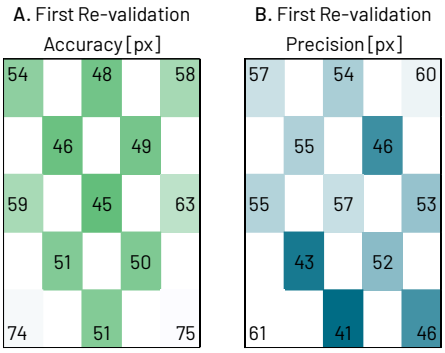

Figure 20. *"Heatmaps" of* ***A****. Accuracy and* ***B****. Precision for specific screen areas in the First Re-validation task on Smartphones.*

When analyzing accuracy on a per-participant-basis rather than per-item, the median accuracy was 61.56 pixels (61.08 pixels without outliers), while the precision was 79.64 pixels (28.36 pixels with outliers removed). The highest accuracy observed was 18.93 pixels, and the average accuracy for the top 30 participants was 36.79 pixels. The average accuracy for the bottom 30 participants was 115.26 pixels, which increased to 101.96 pixels after excluding participants identified as outliers using the MAD method (n = 3). Figure 21 illustrates the distribution of mean accuracies across participants, with red markers indicating the outliers. In the outlier-inclusive sample, 67.62% of participants achieved an accuracy of ≤ 75 pixels, 85.71% achieved ≤ 100 pixels, and 92.38% ≤ 125 pixels. Without outliers, these values increased to 69.61%, 88.24%, and 95.10%, respectively.

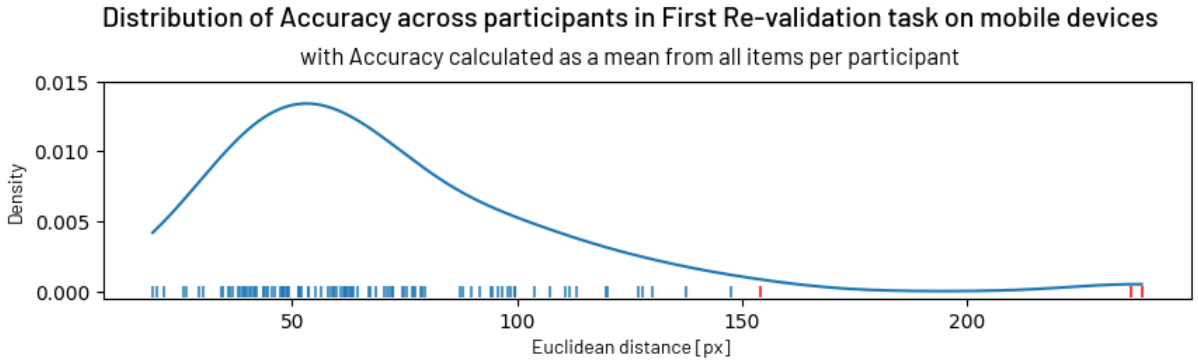

Figure 21. *Distribution of accuracy among participants in the First Re-validation task.*

Second Re-validation

To determine whether the accuracy and precision decreased over time, the metrics from the Second Re-validation were compared to those from the First Re-validation. Table 9 shows the evaluation metrics calculated for each accuracy calculation method for Second Re-validation task.

Table 9. *Evaluation metrics for Second Re-validation task based on different accuracy calculation methodologies on Smartphones.*

Method		Accuracy [px]		Precision [px]	
		outliers included	outliers excluded	outliers included	outliers excluded
(1)	mean_Euclidean_dist_of_all_fixations	111.34	107.16	90.49	72.05
(2)	average_fixation_centroid	104.95	101.03	91.39	73.83
(3)	last_fixation	109.03	107.53	95.45	79.68
(4)	longest_fixation	100.88	99.48	94.75	75.90
(5)	closest_fixation	84.34	83.06	82.05	69.69

For each approach, the accuracy and precision degraded over the course of the experiment. To test if the differences between the two Re-validations were statistically significant, the samples were adjusted to ensure that pairs of data from the same participants were included in both sets, and the Wilcoxon signed-rank test was applied. A statistically significant difference was revealed, with notably lower values observed in the Second Re-validation compared to the First Re-validation (p-value < 0.01). Possible explanations for this could include fatigue from repeated sessions, as the tasks were monotonous and required constant mental and visual concentration.

Full-Screen Accuracy – Large Grid task

To assess the performance of the RealEye webcam eye-tracking system for Smartphones on the wide range of screen areas, the Large Grid task was employed. Since the Large Grid task was not conducted immediately after Calibration, there is a possibility that accuracy results could have improved slightly if it had been, as suggested by the First Re-validation results.

Figure 22 shows the distribution of accuracy across participants for each accuracy calculation method, again revealing a right-skewed pattern for each one of them.

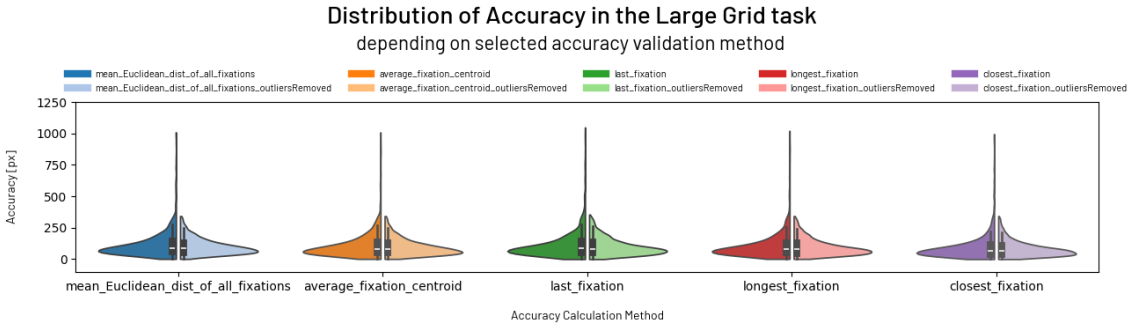

Figure 22 *Distribution of Euclidean distances in the Large Grid task depending on the accuracy calculation method and outliers inclusion/exclusion on Smartphones.*

The results presented in Table 10 show that the accuracy measured using a methodology similar to that proposed by iMotions WebET[12] once again yielded the worst results, with an accuracy of 90.33 pixels (87.97 pixels when outliers were excluded). The best results were obtained using the methodology employed by Nystrom et al.[32], with calibration accuracy reaching up to 70.15 pixels (69.08 pixels when outliers were removed).

Table 10. *Evaluation metrics for First Re-validation task based on different accuracy calculation methodologies.*

Method		Accuracy [px]		Precision [px]	
		outliers included	outliers excluded	outliers included	outliers excluded
(1)	mean_Euclidean_dist_of_all_fixations	90.33	87.97	96.82	62.89
(2)	average_fixation_centroid	83.70	81.03	97.61	64.05
(3)	last_fixation	85.74	84.04	102.83	68.05
(4)	longest_fixation	81.03	79.39	98.76	65.37
(5)	closest_fixation	70.15	69.08	89.00	60.95

In the outlier-inclusive sample, the accuracy calculated for different areas of the screen ranged from 52 to 88 pixels, with the best results obtained for the central and central-left part of the screen, and the worst in the bottom-right corner. Precision was highest in the central part of the screen (middle horizontal column with a minimum of 62 pixels), and lowest in the bottom area (127 pixels). Detailed results for accuracy and precision for different parts of the screen in the Large Grid task are presented in Figure 23.A and 23.B.

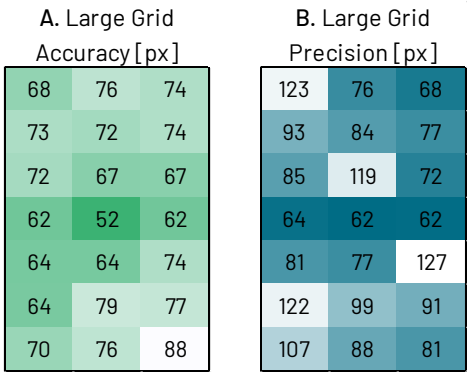

Figure 23. *"Heatmaps" of* **A.** *Accuracy and* **B.** *Precision for specific screen areas in the Large Grid task on Smartphones.*

When analyzing accuracy on a per-participant basis rather than per item, the median accuracy was 69.68 pixels. The highest accuracy observed was 19.94 pixels, and the average accuracy for the top 30 participants was 38.82 pixels. The average accuracy for the bottom 30 participants

was 166.84 pixels, which increased to 115.50 pixels after removing participants identified as outliers using the MAD method (n = 7). Figure 24 illustrates the distribution of median accuracies per participant, with red markers indicating the outliers. In the outlier-inclusive sample, 51.43% of participants achieved an accuracy of ≤ 75 pixels, 76.19% ≤ 100 pixels, and 84.76% ≤ 125 pixels. Without outliers, these values increased to 55.67%, 82.47%, and 91.75%, respectively.

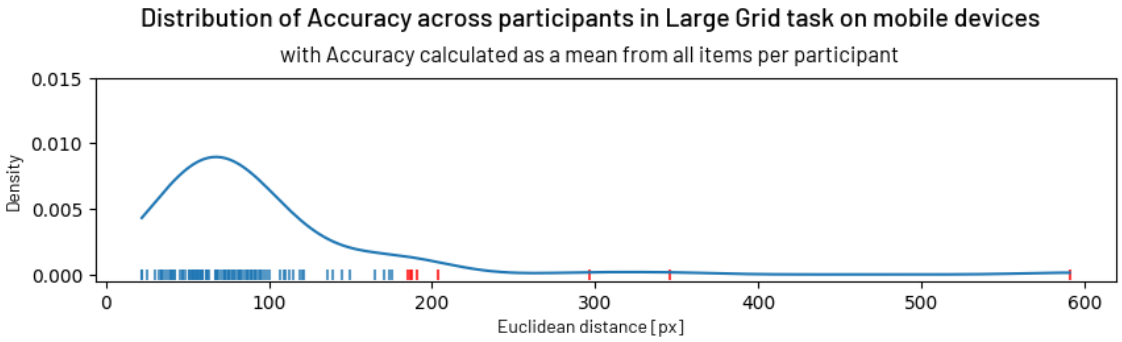

Figure 24. *Distribution of accuracy among participants in the First Re-validation task on Smartphones.*

Survey-based analysis

To examine potential demographic and environmental factors influencing eye-tracking performance, a survey was conducted at the end of the study. It collected data on eye color, use of vision aids, race/ethnicity, and lighting conditions during the experiment. Analyzing these variables helped understand their impact on the overall accuracy and precision of the RealEye system on Smartphones.

Eye color

Participants self-reported their eye color by selecting one of the following options: blue, grey (Bluish), green, hazel (Greenish), amber, or brown (Brownish). Table 11 presents the evaluation metrics for each eye color group in the Large Grid task in the outlier-inclusive sample on Smartphones.

Table 11. *Evaluation metrics for different Eye color groups in the Large Grid task on Smartphones.*

Eye color	Num of participants	Accuracy [px]	Precision [px]
Bluish	42	67.92	62.95
Brownish	38	70.81	66.21
Greenish	25	75.79	138.27

To compare the results among participants, Euclidean distances were analyzed for each eye color group. The distribution of these values is shown in Figure 25. Significant differences were identified using the Kruskal-Wallis test, followed by Dunn's post-hoc test with Bonferroni-adjusted p-values. These differences were observed between the Bluish and Greenish groups, as well as between the Brownish and Greenish groups, with participants with Greenish eyes exhibiting lower accuracy in both cases (p-value < 0.05).

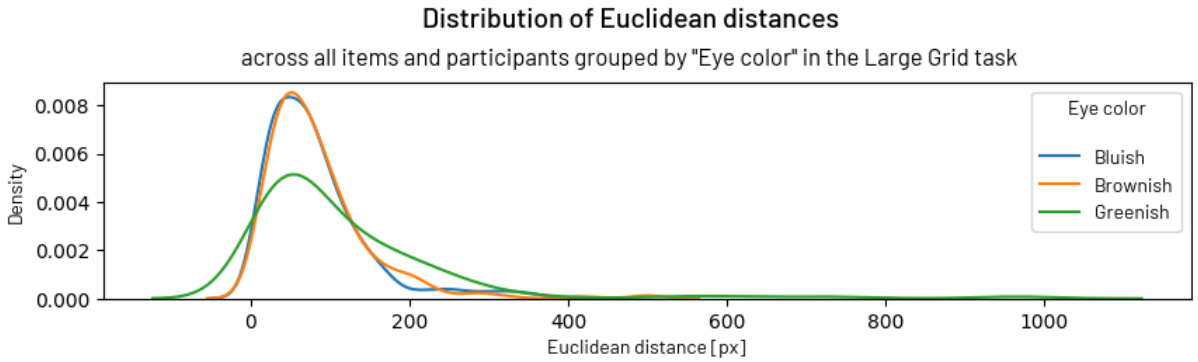

Figure 25. *Distribution of Euclidean distances calculated for specific "Eye color" groups on Smartphones.*

Use of vision aids

Participants reported whether they used vision aids, such as glasses or contact lenses, during the experiment. Table 12 presents the evaluation metrics for each analysis group in the Large Grid task in the outlier-inclusive sample.

Table 12. *Evaluation metrics for different Vision aids groups in the Large Grid task on Smartphones.*

Use of vision aids	Num of participants	Accuracy [px]	Precision [px]
No	64	65.77	60.24
Yes	41	79.01	119.30
Yes, glasses	36	77.06	126.82
Yes, contact lenses	5	82.53	54.91

To compare the results among participants, Euclidean distances were analyzed for each vision aids group. The distribution of these values is shown in Figure 26. The Kruskal-Wallis test revealed significant differences between the groups using ("Yes") and not using ("No") vision aids, particularly between participants who did not use vision aids and those who wore glasses. These findings were confirmed by Dunn's post-hoc test with Bonferroni-adjusted p-values (p-value < 0.01 in both cases). Participants who did not use any vision aids showed significantly better results than those who did, which aligns with

previous research findings[24]. It is worth noting that the sample size for the "Yes, contact lenses" group was n=5, which can be considered small compared to the other groups.

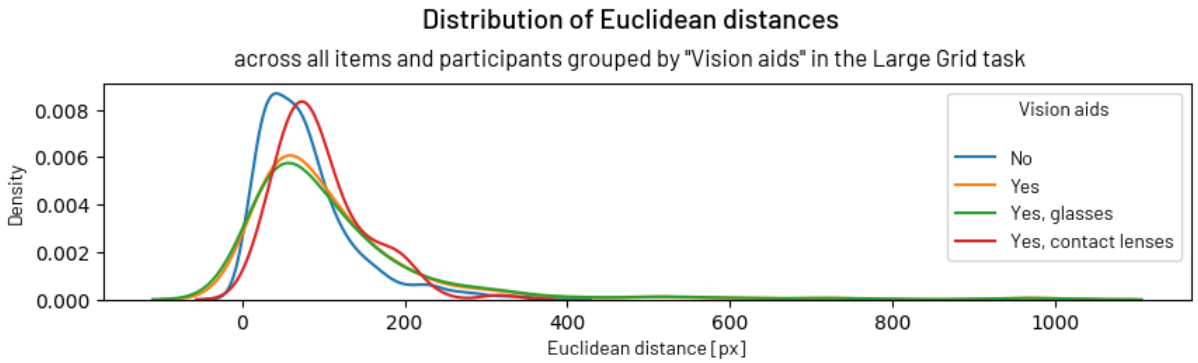

Figure 26. *Distribution of Euclidean distances calculated for specific "Vision aids" groups.*

Race/Ethnicity

Participants were asked to report their race or ethnicity by selecting one of the following: White, Asian, African / African American / Black or Latin American / Hispanic. Table 13 presents the evaluation metrics for each analysis group in the Large Grid task in the outlier-inclusive sample.

Table 13. *Evaluation metrics for different Race / Ethnicity groups in the Large Grid task on Smartphones.*

Race / Ethnicity	Num of participants	Accuracy [px]	Precision [px]
White	92	71.13	89.82
Asian	5	87.30	111.14
African / African American / Black	6	64.68	41.99
Latin American / Hispanic	2	67.04	22.38

Due to little amount of participants in groups of Asian, African / African American / Black, and Latin American / Hispanic, the results of ethnicity groups were not compared between each other.

Lighting conditions

Participants reported their lighting conditions by selecting one of the following options: well-lit room with natural light, well-lit room with artificial light, side lighting, or dimly lit room. Table 14 presents the evaluation metrics for each analysis group in the Large Grid task in the outlier-inclusive sample.

Table 14. *Evaluation metrics for different Lighting conditions groups in the Large Grid task on Smartphones.*

Use of vision aids	Num of participants	Accuracy [px]	Precision [px]
Well-lit room	82	68.12	92.93
Well–lit room with natural light	52	64.83	57.15
Well-lit room with artificial light	30	75.66	131.67
Side lighting	10	88.75	87.23
Dimly-lit room	13	75.49	54.17

Due to little amount of participants in all groups except for the well-lit room with natural and artificial light, the only results compared was between those groups. The Kruskal-Wallis test revealed significant differences between them, with participants in well-lit room with natural light achieving significantly higher accuracy than those in well-lit room with artificial light.

Comparison of RealEye results on Smartphones with other approaches

Due to the absence of a standardized framework and the variability in accuracy calculation methods, a direct comparison with previous studies is not feasible. However, to provide an approximate estimate, accuracy data was converted to centimeters using the Formula 2 below. This conversion was based on a fixed screen resolution of 342×287 pixels and a 4.7-inch diagonal display, as used in iTracker[18] for the iPhone 6s. In the approach used in Sticky by Tobii[13] however, the specific screen resolution(s), screen diagonal(s), and the assumed viewing distance were not explicitly provided. It is important to emphasize that these estimates rely on assumed variables, which could not be directly measured during the study.

Formula 2..*Mathematical equation to convert pixels to centimeters.*

$$Accuracy\ [centimeters] = \frac{Accuracy\ [pixels]}{PPI} \times 2.54$$

Where:

- Pixels Per Inch $PPI = \frac{Diagonal\ Resolution\ [pixels]}{Screen\ diagonal\ [inches]}$, and $Diagonal\ Resolution\ [pixels] = \sqrt{Screen\ width^2 + Screen\ height^2}$

First Re-validation

In the First Re-validation task, the results indicated that 67.62% of participants achieved an accuracy of ≤ 1.32 cm, 85.71% achieved ≤ 1.76 cm, and 92.38% achieved ≤ 2.20 cm. Without outliers, these values increased to 69.61%, 88.24%, and 95.10%, respectively. The average accuracy immediately after calibration = 55.62 px (55.27 px without outliers), is approximately 0.98 cm (0,97 cm).

Large Grid

In the Large Grid task, 51.43% of participants achieved an accuracy of ≤ 1.32 cm, 76.19% ≤ 1.76 cm, and 84.76% ≤ 2.20 cm. Without outliers, these values increased to 55.67%, 82.47%, and 91.75%, respectively. The average "full-screen" accuracy = 70.15px (69.08 px without outliers), is approximately 1.23 cm (1.21 cm).

7. Conclusions and Implications

The study delivered a detailed analysis of the RealEye eye-tracking system for Computers and Smartphones, evaluating its performance across real-life conditions. The results were obtained from different participants in various environments, considering factors like devices used, and lighting conditions, making the results realistic and reliable.

7.1. Desktops and Laptops – key findings

Accuracy and its decay over time

Accuracy ≈ 106 pixels in the First Re-validation task, with the best individual accuracy around 22 pixels. The results indicated that 75% of participants achieved accuracy ≤ 150 pixels, 87% ≤ 200 pixels, and 93% ≤ 250 pixels. However, a significant decline in both accuracy and precision was observed over time with lower results in Second Re-validation task.

Full-screen Accuracy

The Large Grid task evaluated the performance of the system on a wide range of screen areas. **The average "full-screen" accuracy was ~123 pixels**, improving to around 121 pixels with outliers removed. The results showed that 70% of participants achieved an Accuracy ≤ 150 pixels, 79% ≤ 200 pixels, and 84% ≤ 250 pixels. The highest accuracy was observed in the central areas of the screen, and decreasing

progressively towards the edges. It is anticipated that accuracy could be improved if measurements were taken immediately after Calibration.

Survey-Based Analysis

The survey-based analysis provided additional insights into how demographic and environmental factors might influence eye-tracking performance. Significant differences were found between eye color groups – participants with Greenish eyes showed significantly worse results than those with Bluish or Brownish eyes (p-value < 0.05). Moreover, participants who did not use vision aids performed significantly better than those who wore glasses (p-value < 0.01). These findings align with previous research that suggests the use of vision aids can affect eye-tracking accuracy. No differences were observed in performance across different race/ethnicity groups or lighting conditions, due to the limited number of participants in specific subgroups.

7.2. Smartphones – key findings

Accuracy and its decay over time

Average Accuracy ≈ 56 pixels in the First Re-validation task, with the best individual accuracy around 15 pixels. The results indicated that 71% of participants achieved accuracy ≤ 75 pixels, 87% ≤ 100 pixels, and 93% ≤ 125 pixels. However, a significant decline in both accuracy and precision was observed over time with lower results in Second Re-validation task.

Full-screen Accuracy

The Large Grid task evaluated the performance of the system on a wide range of screen areas. **The average "full-screen" accuracy was ~70 pixels**, improving to around 69 pixels with outliers removed. The results showed that 59% of participants achieved an accuracy ≤ 75 pixels, 78% ≤ 100 pixels, and 85% ≤ 125 pixels. The highest accuracy was observed in the central areas of the screen, and decreasing progressively towards the edges. It is anticipated that Accuracy could be improved if measurements were taken immediately after Calibration.

Survey-Based Analysis

The survey-based analysis provided additional insights into how demographic and environmental factors might influence eye-tracking performance. Significant differences were found between eye color groups – participants with Greenish eyes showed significantly worse results than those with Bluish or Brownish eyes (p-value < 0.05). Moreover, participants who did not use vision aids performed significantly better than those who wore glasses ($p < 0.01$). These findings align with previous research that suggests the use of vision aids can affect eye-tracking accuracy. Interestingly, participants in well-lit room with natural light achieved significantly better results that those in well-lit room with artificial light. No differences were observed in performance across different race/ethnicity groups, due to the limited number of participants in specific subgroups.

7.3. Limitations and Implications for future research

The findings from this study provide valuable insights into the performance of the RealEye system while highlighting several areas for future improvement on both Computers and Smartphones. The remote nature of the study, while beneficial for reaching a broad participant base, limited the ability to verify whether participants strictly adhered to instructions that could affect data quality.

Given the observed decay in accuracy over time, future iterations of the RealEye system should explore strategies for minimizing calibration drift or refining recalibration protocols to maintain consistent accuracy during extended use. The decline in accuracy suggests a need for recalibrations or advanced stabilization techniques to uphold performance throughout longer sessions. The monotony of the tasks may have also contributed to a decline in participant engagement and data quality over time. Future studies should consider incorporating more engaging and varied tasks, such as interactive elements that require participants to respond actively.

The variability in accuracy across different screen areas indicates that optimizing the system's algorithms to address performance near the screen edges, could enhance the overall results. At this point, ensuring that areas of interest and key stimuli are located in central regions of the screen, where accuracy is highest, is recommended to achieve optimal performance.

Incorporating personalized calibration options that consider eye color and the use of vision aids, such as glasses or contact lenses, could further improve tracking accuracy and user satisfaction. Future research should also aim to include larger and more diverse samples to better understand the influence of demographic variables on eye-tracking performance. Although lighting conditions had minimal impact in this study, exploring a wider range of lighting environments could provide additional insights into optimizing system performance in varied real-world settings.

The system's validation threshold could be lowered to exclude participants whose data does not meet the necessary quality standards. This adjustment would help ensure that only high-quality, reliable data is included in the analysis, thereby improving the overall accuracy and robustness of the findings.

By addressing these limitations, future research can build upon the findings of this study and contribute to the development of more robust and accurate RealEye webcam eye-tracking systems for computers. The study also highlighted that data quality directly impacts both accuracy and precision, underscoring the importance of continued efforts to refine and enhance the system's performance.

www.realeye.io