Estimating Gaze Duration Error from Eye Tracking Data

John Hawkins
john.hawkins@playgroundxyz.com
Playground XYZ
Sydney, NSW, Australian

ABSTRACT

Eye tracking technologies produce a series of timestamped gaze fixation points that can be attributed to objects within a subject's field of vision. These fixation points are often aggregated in order to generate a measurement of the gaze duration in Area of Interest studies. Error is typically measured on the basis of the variation between individual gaze fixation points and the intended point of fixation. These underlying errors in fixation have no inherent transformation for making a statement about error in gaze duration, even though they are expected to be related.

In this work we develop an algorithm for estimating the error distribution of gaze duration measurement through Monte Carlo simulation using the content of an eye tracking calibration log file. We provide this algorithm in an open source application to allow researchers to understand the error in their gaze duration measurements. We use this application to conduct experiments on the expected error bounds for different duration measurements across a fixed session length for a simulated area of interest study on a mobile device. The results indicate that error in gaze duration estimation is sensitive to fixation error beyond a bound that will depend on the size of the area of interest and the comparative length of the viewing session.

CCS CONCEPTS

• **Mathematics of computing** → **Probabilistic algorithms**; • **Human-centered computing** → **Human computer interaction (HCI)**; • **Information systems** → **Online advertising**.

KEYWORDS

Eye Tracking, Attention Measurement, Digital Advertising, Error Estimation

ACM Reference Format:
John Hawkins. 2023. Estimating Gaze Duration Error from Eye Tracking Data. In *Proceedings of (MLHCI '23).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Eye tracking is an important technology with a wide variety of applications. It permits evaluation of software user interfaces[8], the development of new forms of software interaction, as well as a potential method of biometric identification[15]. Eye tracking has become a core tool for empirical investigations into human behaviour and is widely used as a method of measuring explicit attention to visual stimuli. It has been applied to study psychological phenomena ranging from cognition[3] to mental health[6, 19], and is now routinely used to evaluate advertising [10]. Eye tracking technology has allowed marketing researchers to study many factors that contribute to effective advertising, including brand recall [21], the capture and tranfer of attention [18], the impact of images of faces [5], the attention effects of animation [7], and the relationship with social media posts [1].

Increasingly, eye tracking applications are built using devices that employ machine learning models trained to predict fixation locations from a camera image. While this approach to eye tracking may suffer in terms of accuracy, it can open up possibilities for researchers in terms of reduced cost and increased scale of experiments. As well as facilitating studies that head position restricting devices prevent[20].

The source of errors in eye tracking technology can be decomposed into a range of independent sources that researchers need to consider[11]. Unless they are properly addressed these sources of error can manifest themselves as systematic biases across undesirable dimensions such as the age[4] or ethnicity of subjects[2]. Machine learning based eye tracking solutions can be impacted by the variety of faces and lighting conditions in the training data. The errors in fixation point measurement for a given eye tracking solution are routinely monitored and adjusted through a process called calibration, in which users are presented a series or *required fixation locations* (RFL) which are compared with measured points of fixation[14]. Improvements in the callibration process are an ongoing focus for the development of algorithms[9, 22] and software tools[16]. In some scenarios there are additional considerations such as whether to use fixation data from individual eyes, or a composite signal[13].

In eye tracking literature the error detected in the calibration process is classified using two independent metrics; accuracy and precision. Accuracy is a measure for how close, on average, the fixation points are to the true point of fixation. Precision is the tendency for the measured locations to vary between subsequent measurements for a given point of fixation[12]. This distinction is critical because it forms the basis of methods that can correct for the kind of systematic bias observed with low precision eye tracking. It also been observed that metrics of the precision in these models vary in the way they rank technologies[17]. We can conclude from this that the patterns of systematic error in eye tracking data are more complicated that can be quantifed by a single metric.

When fixation models are used sequentially for the purpose of estimating gaze duration (as is commonly done in point of salience or advertising media studies), then there is the potential for the error

MLHCI '23, March 24–26,2023, Singapore

ACM ISBN 978-1-4503-XXXX-X/23/05…$15.00
https://doi.org/XXXXXXX.XXXXXXX

to either cancel out (reducing error in gaze duration estimation) or to compound (increasing error in gaze duration). Which of these outcomes occurs will depend on the specifics of the model and the conditions under which it is being used. Influencing factors include the ratio of measured gaze duration to the total viewing time and the error profile of the gaze fixation technology.

The use of eye fixation models to estimate gaze duration (or dwell time) in Area of Interest studies (AOI) was discussed by Holmqvist et al. [11]. The authors simulated the impact of gaze fixation error by adding noise according to manufacturer specifications to data sets of low margin areas of interest. We extend this idea to calculate probabilistic bounds on the error in gaze duration using calibration data as a source of noise distribution that is specific to both the study participants and the device/environment of the study. The result is an open source application that may be used by a wide variety of reserachers to provide error bounds on any gaze duration measurements.

2 METHODOLOGY

We estimate the gaze duration measurement error through a Monte Carlo simulation using the eye tracking model's calibration file of fixation errors. The error profile of the eye tracking model allows us to generate sequences of true fixations and simulate measured fixations with the observed error of the model. These sequences of true and measured fixation are then converted into a distribution of expected gaze duration on a specifc area of interest, for a given measurement.

The algorithm consists of two core steps, first estimating the distrubtion of measurements for a given true duration. Secondly, inverting this into the distribution of true gaze durations for a given measurement. Note, that the algorithm requires that the area of interest be defined as a fixed bounding box and that the distributions are estimated for a fixed session length. This session length is the total period of time that the area of interest was in a subject's field of view, and thus represents the maximum possible gaze duration.

2.1 Measurement Distribution

In the first stage of the methodology we provide an estimate of the distribution of measured durations for all possible true gaze durations. These distributions are produced by generating random fixation paths across the available screen dimensions, and adding noise that is drawn from the calibration data to make it consistent with the observed properties of the gaze model. This produces a set of distributions that capture the expected variation in measurement depending on the length of true fixation.

Formally stated this is the probability distribution over measured durations $\hat{d}$ given a known duration d, the session length s and the gaze target location l. We express this distribution as $P(\hat{d}|d, l, s)$. The procedure for making this estimation involves iteratively generating random samples of gaze fixation paths $\mathcal{F}$ consistent with the parameters d, s and l. For each point f in $\mathcal{F}$ we determine a measurement point $\hat{f}$ by drawing samples from the eye tracking calibration set.

The samples from the calibration data are drawn as a delta in the two dimensional Cartesian space of the field of view. Each delta consists of δ_x and δ_y which correspond to the error in the x and y dimensions. We apply these delta values to the true gaze fixation point f to create the simulated measurement $\hat{f}$. We draw these delta values from the calibration file such the the probability of choosing a delta is propotional to the distance between the point f in the generated path, and the true fixation point in the callibration set.

The complete algorithm for estimating $P(\hat{d}|d, l, s)$ is shown in Algorithm 1

Algorithm 1 Estimation of $P(\hat{d}|d, l, s)$

Input: l, s, C
Output: $P(\hat{d}|d, l, s)$
  $N \leftarrow samples$ ▷ Configure simulation samples per duration
  $D \leftarrow floor(s/increment)$ ▷ Durations are sampled at discrete intervals
  $P \leftarrow Array(D+1, D+1)$ ▷ The distribution P is a 2 dimensional array
  for $d \in 0toD$ **do** ▷ Iterate over all possible true gaze durations
    for $n \in 0toN$ **do** ▷ Iterature to collect the N samples for d
      $p \leftarrow generatePath(d, l, s)$
      $d_e \leftarrow generateMeasurement(p, l, C)$
      $P[d, d_e] += 1/N$ ▷ Probability increment for a measuremnt of d_e
    end for
  end for
  return P

The bulk of the work is done by two functions inside the simulation loop. The first function $generatePath(d, l, s)$ takes a true gaze duration (for that simulation iteration) the location of the gaze target and the viewing session length. It then generates a random path of fixation points that is compatible with those parameters. The second function $generateMeasurement(p, l, C)$, takes the path, target location and the eye tracking calibration data and produces a sample of measured attention time. This is achieved by using the calibration data to sample fixation errors in the proximity of each location within the path and applying them to the path data. The noisy path data is then used to determine the estimated gaze duration by looking at the number of fixation points that intersect with the target location. Repeated application of this process for a given value of d produces multiple samples of $\hat{d}$ for that true gaze duration. Repeating the process over different values of d allows us to populate a two dimensional array where the first dimension is the true duration and the second the estimated duration.

Note that our probability distribution is discrete because we restrict ourselves to sampling over a set of fixed time period increments (i) between zero and the session length. This restriction is realistic for most machine learning eye tracking systems as they output fixation points at fixed intervals of time.

2.2 Gaze Duration Distribution

In real world applications we will have the true session length but rely on the model for the estimate of gaze duration. Meaning that the probability distribution we want is the distribution over true gaze durations for a given measurement. In the previous section

we estimated the inverse of this: the distribution of measurements of a given true duration.

In order to estimate the desired distribution $P(d|\hat{d}, l, s)$ we need to use Bayes' rule, as shown in Equation 2.1.

$$P(d|\hat{d}, l, s) = \frac{P(\hat{d}|l, s, d)P(d|l, s)}{P(\hat{d}|l, s)} \tag{2.1}$$

We use a uniform prior for $P(d|l, s)$, meaning in the absence of additional information all gaze durations less than the session length are equally likely. As our distributions are discrete estimations of an underlying continuous distribution the value of each discrete value in $P(d|l, s)$ is equal to $1/(1 + floor(s/i))$. We can estimate the value of the denominator $P(\hat{d}|l, s)$ by iterating over all possible values of $\delta \in D$ and summing the product of $P(\hat{d}|l, s, \delta)$ and $P(\delta|l, s)$. This fully explicated form is shown in Equation 2.2.

$$P(d|\hat{d}, l, s) = \frac{P(\hat{d}|l, s, d)P(d|l, s)}{\sum_{\delta \in D} P(\hat{d}|l, s, \delta)P(\delta|l, s)} \tag{2.2}$$

2.3 Implementation

The algorithm described in the preceding sections has been implemented as an open source python package called *gazerr*. It can be used as a software library, exposing functions for calculating the probability densities, or it can deployed as a command line application. The source code is available on GitHub[1] and the python package can be installed from PyPi.[2]

The source repository for the *gazerr* application also contains the complete set of scripts for re-creating the data sets and running the experiments outlined in the next section. Please refer to the README file for the sequence of steps required for replication of all results.

3 EXPERIMENTS

We apply the *gazerr* application to investigate the relationship between measured gaze duration and expected true duration of fixation under a variety of simulated scenarios. Each scenario involves a variation in the underlying error indicated by the eye tracking calibration file and length of the session in which the user could be looking at the area of interest.

For the sake of our experiments we use a simulated device, approximately the size of a small smart-phone, with viewport pixel dimensions of 350 by 627. Our area of interest (AOI) is a fixed position rectangle that resembles an iAB standard medium rectangle ad unit (MREC). We position the simulated ad toward the top of the screen. We iterate over a range of potential mean error values from 20 to 160 pixels (equivalent to varying the accuracy of the eye tracking models). These mean error values determine the average Euclidean distance between a measured fixation point and the true fixation point.

For each of these mean error values we can generate synthetic eye tracking calibration files containing a sample of true fixation points and measured fixation points that are coherent with the simulation parameters. We produce three versions of this calibration data to simulate and experiment with multiple real-world scenarios:

- `Unbiased`: Error distributed uniformly around true fixation point.
- `Biased`: Error tending toward upper left of true fixation.
- `Biased Precise`: Error tending toward upper left but clustered.

These three scenarios are listed above. The unbiased calibration data is distributed evenly and centered around each of the true target positions in the callibration file. The biased data imposes a tendency on the error to be pushed toward the upper left of the screen away from the true fixation position. The biased precise data is again forced toward the upper left corner of the screen, but constrained so that measurements cluster together, meaning that although measurements have the same mean error that are more tightly clustered with each other around the mean error position.

We include the two biased calibration datasets as they are consistent with both our observations of real eye tracking data and what is found in the literature[12]. We use this synthetic data creation in order to explore the extent to which different patterns in error, bias and precision will affects gaze duration measurements.

We feed each of these synthetic calibration files into the *gazerr* application to calculate the posterior probability distributions over true gaze duration for the set of measured gaze durations. These posterior distributions are converted into expected values of true gaze duration for each measured duration.

In these experiments we use both a fixed session length of one thousand milliseconds and a fixed position for the area of interest. These restrictions allow us to focus on the impacts of mean error, bias and precision on the quality of gaze duration measurment.

4 RESULTS

The first experimental result is the expected true duration against the measured duration for a single eye tracking application with a mean fixation error of 70 pixels. These results are shown in Figure 1.

The result shown in Figure 1 is indicative of what we see under multiple variations of the simulation configuration. At lower meaurement values there is a tendency for the measurement to be an accurate estimation of true gaze duration, while at higher values it tends toward an under-estimation of the true gaze duration. This problem is worse for the model with the biased error, but it is mitigated when the biased error comes from a model with higher precision. This is potentially due to the fact that the high prescion scenario has lower tendency for gaze to be pushed off screen.

We next examined the way that exected error in the gaze duration estimate tracks with expected error in the gaze fixation points. This involved calculating the mean absolute error in gaze duration for each of the mean fixation error validation files. We performed this calculation using two methods. The first method, labelled uniform, assumes that all measured values are equally likely. The second method, labelled exponential, assumes that low measurements are much more likely than higher measurement, and that this drop in likelihood follows an approximately exponential distribution.

As shown in Figure 2 we see that the error in gaze duration grows with error in fixation points, as would be expected. However, when we use the exponential weighting to calculate the expected error in duration the gaze duration error grows much slower than in

[1] https://github.com/playground-xyz/gazerr
[2] https://pypi.org/project/gazerr/

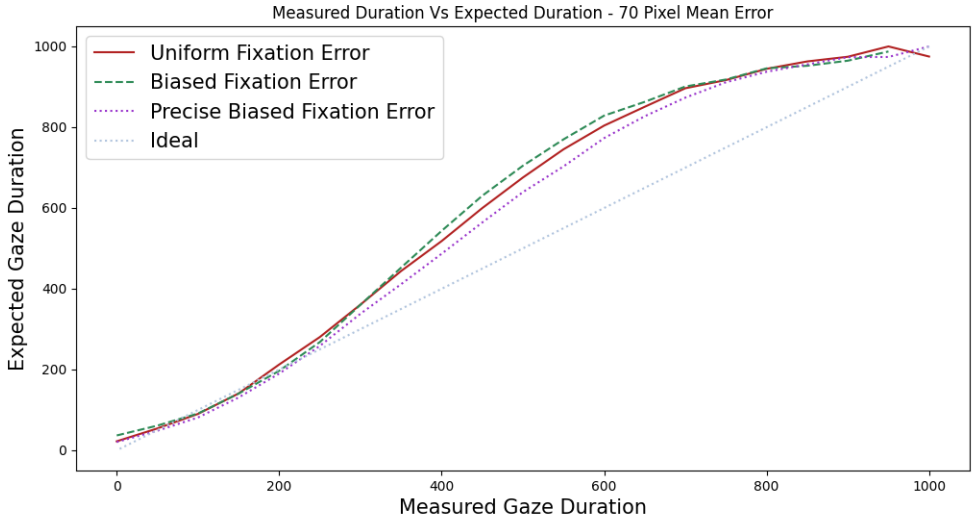


Figure 1: Measured Gaze Duration Versus Expected Gaze Duration. MREC ad appearing on screen for 1s with 70px Mean Error in Fixation

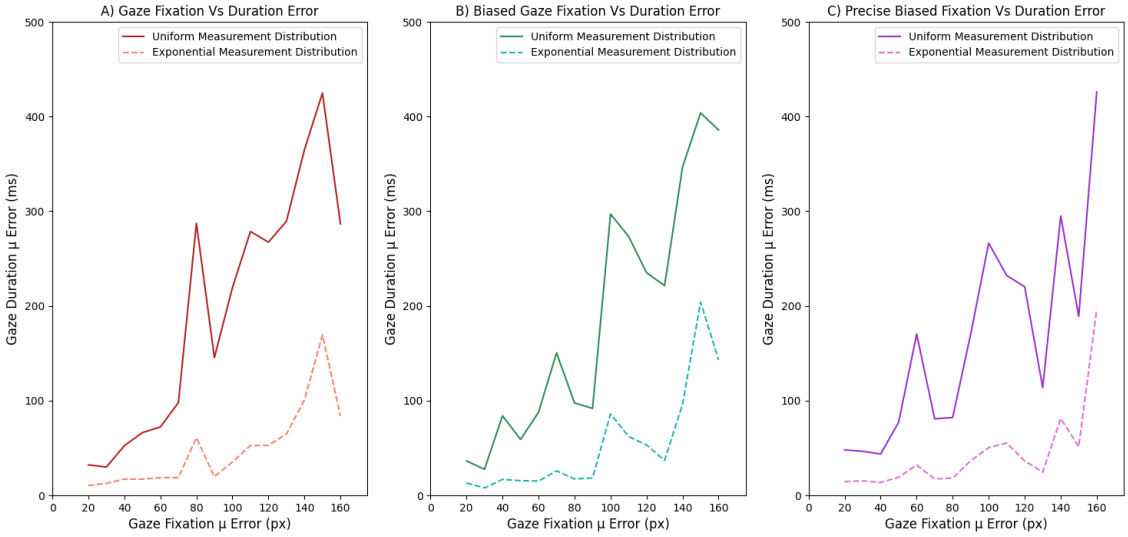


Figure 2: Fixation error versus gaze duration error.

the uniform case. This suggests that the observed tendency toward smaller measurements in advertising area of interest studies means lower expected error overall. This pattern is true for the centrally distributed error (A) as well as the biased calibration error (B) and precise biased calibration error (C) scenarios.

In addition, the additional error in gaze duration we see when the error is biased in a particular direction (shown in B) is partly mitigated by a fixation model that is more precise (shown in C). This suggests that a precise fixation sampling process is able to partly overcome the false positive and false negatives of an error prone gaze fixation model.

In our final experiment we look at the effect of both the size of the measured gaze duration and the mean error of fixation on the expected error in gaze duration. We conduct this experiment using only the unbiased synthetic data in order to understand how error centred on true fixation points relates to error in gaze duration

across the range of measurement. We display these results as a three dimensional expected error surface over these two key dimensions in Figure 3

The left hand plot shows the complete error surface in which you will observe large spikes in error in the corner corresponding to both high mean error for the fixation models and high measurements of gaze duration. This effect is caused by a complete absence of probability density in these regions, meaning that the error is maximal because the expected value is zero. This effect was consistently observed for eye tracking models with high mean fixation error and when the measured duration was high. The consequence of this is that our model estimates that when fixation error is this high, measurements of these values become extremely unlikley to be observed.

We added the right hand plot in Figure 3 to illustrate the error surface in the absence of these deceptive outliers. We see that, in general, the expected error in a gaze duration measurement is maximal around the mid point of the measurement range. The effect is lessened by lower mean error in the fixation models, but remains consistent for all levels of fixation error.

5 CONCLUSION

We have demonstrated that the statistial information contained in eye tracking calibration data can be utilised for estimating errors in gaze duration estimation. The process involves simulating the distribution of measured duration for each possible true duration, and then inverting it with Bayes rule into a distribution over true duration for a given measurement.

We have released this approach as the open source application *gazerr* and used it to explore the relationship between fixation error and duration error, under range of realistic error profiles when looking at measuring attention on a standard size ad within a mobile screen. Our results show that the nature of the error in gaze duration tends to depend on the size of the measurement. Smaller measurements tend to have lower expected error while larger measurements tend toward under prediction. We plotted the relationship between expected error and saw that the observed tendency toward smaller duration measurements (estimated with an psuedo-exponential distribution) delivers a slower growing expected duration error with the mean error in fixation points.

Finally, our simulations revealed that for large error in fixation points, certain measurements become so unlikely that our Monte Carlo simulations placed none of the probability density in those regions. This strongly suggests that an additional downside to large fixation point error is a reduced range of measurement for area of interest studies.

REFERENCES

[1] Ana Barreto. 2013. Do users look at banner ads on Facebook? *Journal of Research in Interactive Marketing* 7 (05 2013). https://doi.org/10.1108/JRIM-Mar-2012-0013

[2] Pieter Blignaut and Daniël Wium. 2013. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior research methods* 46 (04 2013). https://doi.org/10.3758/s13428-013-0343-0

[3] Tad Brunyé, Trafton Drew, Donald Weaver, and Joann Elmore. 2019. A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications* 4 (12 2019). https://doi.org/10.1186/s41235-019-0159-2

[4] Kirsten Dalrymple, Marie Manner, Katherine Harmelink, Elayne Teska, and Jed Elison. 2018. An Examination of Recording Accuracy and Precision From Eye Tracking Data From Toddlerhood to Adulthood. *Frontiers in Psychology* 9 (05 2018), 803. https://doi.org/10.3389/fpsyg.2018.00803

[5] Soussan Djamasbi, Marisa Siegel, Thomas Tullis, and Rui Dai. 2010. Efficiency, Trust, and Visual Appeal: Usability Testing through Eye Tracking. In *In System Sciences (HICSS), 2010 43rd Hawaii International Conference*. 1 – 10. https://doi.org/10.1109/HICSS.2010.171

[6] Almudena Duque and Carmelo Vazquez. 2014. Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study. *Journal of Behavior Therapy and Experimental Psychiatry* 46 (09 2014). https://doi.org/10.1016/j.jbtep.2014.09.005

[7] Kai-Christoph Hamborg, M. Bruns, Frank Ollermann, and Kai Kaspar. 2012. The effect of banner animation on fixation behavior and recall performance in search tasks. *Computers in Human Behavior* 28 (03 2012), 576–582. https://doi.org/10.1016/j.chb.2011.11.003

[8] Katarzyna Harezlak, Jacek Rzeszutek, and Pawel Kasprowski. 2015. The Eye Tracking Methods in User Interfaces Assessment. In *Intelligent Decision Technologies*, Rui Neves-Silva, Lakhmi C. Jain, and Robert J. Howlett (Eds.). Springer International Publishing, Cham, 325–335.

[9] Almoctar Hassoumi, Vsevolod Peysakhovich, and Christophe Hurter. 2019. Improving eye-tracking calibration accuracy using symbolic regression. *PLOS ONE* 14 (03 2019), e0213675. https://doi.org/10.1371/journal.pone.0213675

[10] Guillaume Hervet, Katherine GuÃŠrard, SÃŠbastien Tremblay, and Mohamed Chtourou. 2011. Is Banner Blindness Genuine? Eye Tracking Internet Text Advertising. *Applied Cognitive Psychology* 25 (09 2011), 708 – 716. https://doi.org/10.1002/acp.1742

[11] Kenneth Holmqvist, Marcus NystrÃśm, and Fiona Mulvey. 2012. Eye tracker data quality: What it is and how to measure it. *Eye Tracking Research and Applications Symposium (ETRA)* (03 2012). https://doi.org/10.1145/2168556.2168563

[12] Kenneth Holmqvist, Saga Lee Örbom, Ignace TG. C. Hooge, Diederick C. Niehorster, Robert G. Alexander, Richard Andersson, Jeroen S. Benjamins, Pieter Blignaut, Anne-Marie Brouwer, Lewis L. Chuang, Kirsten A. Dalrymple, Denis Drieghe, Matt J. Dunn, Ulrich Ettinger, Susann Fiedler, Tom Foulsham, Jos N. van der Geest, Dan Witzner Hansen, Samuel B. Hutton, Enkelejda Kasneci, Alan Kingstone, Paul C. Knox, Ellen M. Kok, Helena Lee, Joy Yeonjoo Lee, Jukka M. Leppänen, Stephen Macknik, Päivi Majaranta, Susana Martinez-Conde, Antje Nuthmann, Marcus Nyström, Jacob L. Orquin, Jorge Otero-Millan, Soon Young Park, Stanislav Popelka, Frank Proudlock, Frank Renkewitz, Austin Roorda, Michael Schulte-Mecklenbeck, Bonita Sharif, Frederick Shic, Mark Shovman, Mervyn G. Thomas, Ward Venrooij, Raimondas Zemblys, and Roy S. Hessels. 2022. Eye tracking: empirical foundations for a minimal reporting guideline. *Behavior Research Methods* (06 Apr 2022).

[13] Ignace T. C. Hooge, Gijs A. Holleman, Nina C. Haukes, and Roy S. Hessels. 2019. Gaze tracking accuracy in humans: One eye is sometimes better than two. *Behavior Research Methods* 51 (2019), 2712–2721. Issue 6.

[14] Anthony J. Hornof and Tim Halverson. 2002. Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, & Computers* 34, 4 (01 Nov 2002), 592–604.

[15] Paweł Kasprowski and Katarzyna Harezlak. 2018. Biometric Identification Using Gaze and Mouse Dynamics During Game Playing. In *Beyond Databases, Architectures and Structures. Facing the Challenges of Data Proliferation and Growing Variety*, Stanisław Kozielski, Dariusz Mrozek, Paweł Kasprowski, Bożena Małysiak-Mrozek, and Daniel Kostrzewa (Eds.). Springer International Publishing, Cham, 494–504.

[16] Pawel Kasprowski and Katarzyna Harezlak. 2018. ETCAL - A versatile and extendable library for eye tracker calibration. *SoftwareX* 8 (03 2018). https://doi.org/10.1016/j.softx.2017.12.005

[17] Diederick Niehorster, Raimondas Zemblys, Tanya Beelders, and Kenneth Holmqvist. 2020. Characterizing gaze position signals and synthesizing noise during fixations in eye-tracking data. *Behavior Research Methods* 52 (05 2020). https://doi.org/10.3758/s13428-020-01400-9

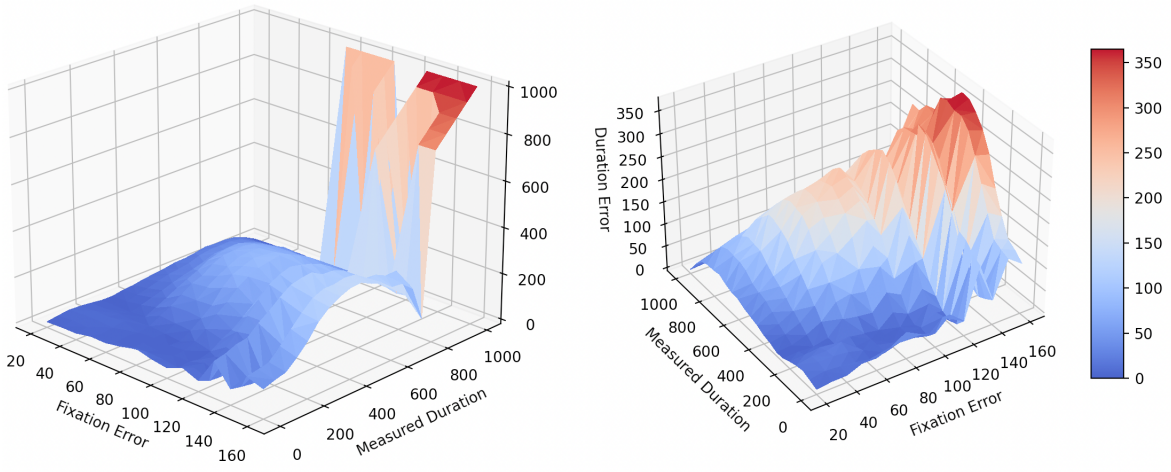
[18] Rik Pieters and Michel Wedel. 2004. Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects. *Journal of Marketing* 68 (05 2004), 36–50. https://doi.org/10.1509/jmkg.68.2.36.27794

[19] Anat Rudich-Strassler, Nimrod Hertz-Palmor, and Amit Lazarov. 2022. Looks interesting: Attention allocation in depression when using a news website - An eye tracking study. *Journal of Affective Disorders* 304 (02 2022). https://doi.org/10.1016/j.jad.2022.02.058

[20] Niilo V. Valtakari, Ignace T. C. Hooge, Charlotte Viktorsson, Pär Nyström, Terje Falck-Ytter, and Roy S. Hessels. 2021. Eye tracking in human interaction: Possibilities and limitations. *Behavior Research Methods* 53, 4 (01 Aug 2021), 1592–1608.

[21] Michel Wedel and Rik Pieters. 2000. Eye Fixations on Advertisements and Memory for Brands: A Model and Findings. *Marketing Science* 19 (11 2000), 297–312. https://doi.org/10.1287/mksc.19.4.297.11794

[22] Yunfeng Zhang and Anthony Hornof. 2014. Easy post-hoc spatial recalibration of eye tracking data. In *Eye Tracking Research and Applications Symposium (ETRA)*. 95–98. https://doi.org/10.1145/2578153.2578166



A) Error Surface - All Measurements

B) Error Surface - Excluding No Density Regions

Figure 3: Mean gaze duration error over measured error and fixation MAE.