Webcams can be used to study eye movements during reading

Marina Serrano-Carot[1], Bernhard Angele[1], Hemu Xu[2], Martin R. Vasilev[2 *]

1. Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Spain

2. Department of Experimental Psychology, University College London, UK

* Corresponding author:

Martin R. Vasilev (m.vasilev@ucl.ac.uk)

Department of Experimental Psychology

26 Bedford Way,

London, WC1H 0AP

United Kingdom

Abstract

Eye-movements have revealed a wealth of information about how readers process written language. Modern eye trackers can record eye-movements in the lab with high accuracy. However, such equipment is expensive, and participant testing is labour-intensive. Kaduk et al. (2023) recently reported that webcam eye-tracking may be a viable alternative, showing that webcams achieve ~0.83 correlations with laboratory eye-trackers during a free-viewing task. We conducted three experiments to test whether webcams can be used to study eye-movements during reading. Participants read short sentences with a target word frequency manipulation while their eye-movements were recorded with both an Eyelink and a webcam in the laboratory (co-registration, Experiments 1 and 2) or online with their own webcam (online study, Experiment 3). In the lab, the webcam-based system recorded raw gaze positions that were highly correlated with the Eyelink (median r= 0.805), replicating Kaduk et al.'s results. We also found evidence for word frequency effects in the webcam data, but there was more noise than in the Eyelink data. In the online study, data quality was much poorer, but we still found reliable effects that were consistent with those observed in the laboratory. In summary, webcam-based eye-tracking may be viable for studying reading, especially under laboratory conditions. However, obtaining useful data may require collecting many more observations when using webcams compared to high-accuracy eye-trackers.

Keywords: eye-movements, reading, eye-tracking, webcameras, online testing

Learning to read and write was a fundamental achievement in human history which paved the way for the development of a knowledge-based modern society. The ability to decode information that others have written, along with the ability to encode new written information, underlies almost every area of human activity today. Much of what we know about language processing during reading comes from the study of eye movements. The importance of eye movements for vision was clear even to ancient scholars (Wade, 2010), but, until the invention of eye-trackers, there was no method to study eye movements beyond introspection and direct observation of a participant while they are reading. Introspection is unreliable for many reasons, but, in the case of eye movements, it is particularly unsuited because saccadic suppression makes it impossible to directly perceive one's own eye movements (Matin, 1974). Direct observation of eye movements in other people does not have this issue, but it is very difficult to do it accurately without technical instruments.

Huey (1908) described some of the early attempts made by researchers such as Javal, Dodge, and Erdmann to record eye movements using devices attached to the eyeball and photographic methods (see also Wade & Tatler, 2008). Researchers like Yarbus (1967; see also Tatler et al., 2010; Richardson & Spivey, 2008) explored these new eye-tracking technologies in order to characterise eye movements in a variety of tasks, such as scene viewing. However, it was not until the introduction of infrared-based eye-trackers in the 1970s that eye movements emerged as a method of observing the moment-to-moment evolution of language processing during reading. The use of infrared light allowed for a much better illumination of the eyes compared to visible light and highly sensitive photodiodes enabled researchers to track the relative positions of the light reflections on the outer and

inner parts of the eye with high precision. Such Dual Purkinje Image (DPI) eye-trackers (Cornsweet & Crane, 1973) allowed researchers to record both the temporal and spatial characteristics of eye-movements with high precision and made it possible to calculate word-based measures, such as average fixation duration on a target word (e.g., Ehrlich & Rayner, 1981). Moreover, the use of DPI eye trackers allowed researchers to change the stimuli presented on the computer display dependent on the participants' gaze position (gaze-contingent display changes). This led to the creation of novel experimental techniques, such as the moving window paradigm, where all text outside of a certain area around the fixation position is masked (McConkie & Rayner, 1975) and the boundary paradigm, where the parafoveal preview that readers receive of one specific target word is manipulated before readers fixate the target word directly. Once a reader's gaze crosses the invisible boundary between the pre-target word and the target word, the display is changed to show the actual target word, making it possible to manipulate the parafoveal information that a reader receives about a word without the participant being consciously aware of the manipulation (Rayner, 1975). Modern video-based infrared eye-trackers, such as SR Research's Eyelink 1000+, have replaced the photodiodes with high-speed video cameras which have much higher tolerances for participant head movements. Using these and other eye-tracking techniques, we have learned much about how readers process words and sentences (for a review, see Rayner, 1998, 2009; as well as Schotter et al., 2012; and Clifton et al., 2016).

Eye-tracking research using video-based infrared eye-trackers has many advantages: It is non-invasive, the methodology is relatively easy to learn and works on most participants, and it can be set up in any laboratory space that is away from direct sunlight, has enough space

for participants to be seated at the recommended monitor and camera distances, and has an electrical outlet. Infrared-based eye-tracking can also be combined with other methods such as EEG (Dimigen et al., 2011; see also Degno et al., 2021), MEG (Pan et al., 2021), and fMRI (Choi et al., 2014). However, as Angele and Duñabeitia (2024) showed, the vast majority of eye-tracking research in reading is done in only a handful of countries, mainly in North America, Western Europe, and, most recently, China. They argue that this is likely due to the cost of a modern, high-precision eye-tracker, which, although moderately priced compared to other methods such as fMRI, MEG, and EEG, still represents a substantial investment for many researchers in countries where research funding is difficult to obtain. Because of this, many languages and writing systems are understudied or not studied at all using eye-movements.

One possible solution for this issue comes from online research. During the COVID-19 pandemic, very little eye-tracking research could be done in the laboratory, but many researchers started doing behavioural language research online, such as studies using the lexical decision task. Paid platforms such as Psychopy/Pavlovia (Bridges et al., 2020; Peirce et al., 2022), Gorilla (Anwyl-Irvine, Dalmaijer, et al., 2020; Anwyl-Irvine, Massonnié, et al., 2020), and open-source solutions such as JATOS (Lange et al., 2015) or lab.js (Henninger et al., 2022) take advantage of the performance of Javascript in the browser in order to provide reasonably high-precision and accuracy stimulus presentation and response collection on virtually any PC or, in some cases, mobile device. For example, in an online masked priming experiment, Angele et al. (2022) found priming effects for stimulus presentation times as low as 33 ms and 50 ms, while even shorter stimulus presentation times did not result in priming

effects (this absence of a priming effect had been previously observed in lab-based studies). This suggests that even experiments that depend on very accurate stimulus timing can be performed online and yield results equivalent to a laboratory study. This not only enabled researchers to continue collecting data during the pandemic but also opened the possibility of reaching participants who would not easily be able to go to the laboratory.

While there are some paradigms to study reading without recording eye-movements, such as self-paced reading or the maze task (Forster et al., 2009), the reaction times from those tasks cannot provide an "online" account of ongoing processing in the same way that eye-movements can. Is it possible then to track eye-movements in an online experiment? In principle, most mobile devices and laptop computers today are equipped with front-facing webcams, which could be used to monitor a participant's eye position. In practice, calculating the gaze position on the screen from a webcam image faces a number of challenges: First, webcams are optimised for recording an acceptable image of the user for videocalls, using visible light. There is no dedicated infrared illuminator such that any recording needs to rely on the illumination produced by the screen and the ambient lighting. In an online experiment, we can control what is on the screen, but we have no control over the lighting environment. We can ask the participant to move to a well-lit area and avoid being back-lit, but our ability to control whether the participant is following these instructions is limited. This alone makes identifying features such as the eyes, the pupil, or the corneal reflection much more difficult. In fact, since there is no dedicated illuminator, there may not be a consistent corneal reflection to track, so a webcam-based eye-tracking method likely has to rely on the orientation of the pupil in relation to the rest of the face to identify gaze position.

Additionally, both the spatial (usually 720p or 1280x720) and the temporal resolution (30 frames per second, fps) of typical webcams is quite low. With built-in webcams, there is the additional issue that these devices are usually optimised to be very small and thin, requiring the use of small, single-lens setups and small sensors that limit image quality. In terms of the spatial resolution, it is therefore clear that the webcam eye-tracking data will be much less precise and accurate than the data recorded by a laboratory-based eye tracker. To compensate for this, an online eye-tracking application will therefore have to apply some data processing, often using machine learning algorithms, to arrive at an acceptable data quality. To maintain participant privacy and avoid having to upload and store videos of the participant's face, this post-processing must be done locally (and in the browser), limiting the computational resources available.

In terms of the temporal resolution, or sampling rate, there is the concern that it may not be possible to accurately detect the main components of eye-movements during reading: saccades, which are fast, ballistic eye movements that move the gaze from one part of the visual scene to another, and fixations, which are the periods of relative gaze position stability in between saccades during which visual information about the text is obtained. Based on the Nyquist-Shannon sampling theorem (Shannon, 1949), a sampling rate of 148 Hz would be required to capture all the information present in the gaze velocity signal, which then is used to detect saccades and fixations (Andersson et al., 2010; Angele et al., 2025). A sampling rate of 30 fps (or Hz) is significantly lower than this. However, even if some velocity information is lost due to a low sampling rate, the resulting detected fixation durations might still contain useful information about the reading process. Angele et al. (2025) simulated low sampling

rates by eliminating samples from data collected with an Eyelink Portable Duo at sampling rates between 250 and 2000 Hz. Even at simulated sampling rates as low as 31.25 Hz, they were able to replicate the word frequency effect on fixation times in reading, i.e., the observation that a word that occurs in the language frequently is, on average, fixated for less time than a word that occurs very rarely (Rayner & Duffy, 1986). The word frequency effect is considered a "benchmark" effect in the study of reading, as it has been found reliably in a large number of studies (Rayner, 1998) and is considered evidence of the impact of cognitive processes (learning) on eye movements (Brysbaert et al., 2018). This demonstrates that even eye movements recorded at only around 30 Hz can contain useful information about ongoing cognitive processing. In principle, this should also be true for eye movements recorded with a webcam.

Recent work has further tested this assumption in applied contexts. For example, van Boxtel et al. (2024) compared WebGazer-based online eye tracking (≈ 14 Hz sampling) with laboratory recordings (≈ 430 Hz) in individuals with aphasia and control participants. They found that webcam tracking reliably captured major group-level differences in sentence comprehension, though with reduced temporal precision and greater variability. This provides important empirical support that even very low–sampling-rate webcam recordings could capture meaningful psycholinguistic effects. Complementary evidence comes from Hutt et al. (2023), who used WebGazer to record gaze data during online reading and found that webcam-based eye-movement features predicted comprehension errors and episodes of mind wandering. These findings show further promise in using low-sampling-rate webcam eye tracking to study natural reading.

Since it seems that the temporal resolution of webcams may, at least in theory, be sufficient for eye-tracking during reading, the question is whether the quality of the spatial data obtained by webcam eye-tracking is good enough to provide useful information for reading research. Spatial data quality can be quantified in two ways: Accuracy is the difference between the recorded gaze position and the true gaze position, while precision indicates whether the eye-tracker gives consistent measurements. As screen sizes and viewing distances vary, accuracy is usually given as the average error in degrees of visual angle and precision is given as the root mean square of the sample-to-sample distances (S2S RMS), again in degrees of visual angle. For example, Angele et al. (2025) found that the average error for the Eyelink Portable Duo (a high precision, laboratory-based eye tracker) was between .44 and .50 degrees of visual angle (using calibration points as the ground truth) and the S2S RMS was between .50 (at 250 Hz) and .11 degrees of visual angle (at 2000 Hz).

In the last ten years, a few online, browser-based eye-tracking software packages have become available. The first of these is WebGazer (Papoutsaki et al., 2016), which uses open-source Javascript algorithm to detect face, eyes, and eye features (such as pupil shape) in real time and constantly self-calibrates by assuming that the user is looking at the location of any mouse movements and clicks made. WebGazer is relatively straightforward to deploy by adding a few lines to the code of any website and, due to the relative computational simplicity, works on most devices that have a modern browser supporting Javascript, a webcam, and mouse or touch input. The authors claim that they achieve sufficient accuracy to obtain useful gaze position data, with an average error of 4.17º of visual angle compared to a laboratory-based eye tracker (Tobii EyeX). They give no information on precision.

Even though this error seems quite large compared to a laboratory-based eye-tracker, there are some very encouraging results for using WebGazer in the visual world paradigm.

The visual world paradigm involves presenting participants with a display containing a number of objects (the "visual world") while presenting a spoken sentence, without giving them any instructions except to listen to the sentence and look at the scene. Tracking participants' eye-movements in this task reveals that they tend to look at any object that is mentioned in the spoken sentence and that these gaze-shifts occur as soon as it becomes clear that an object will be mentioned. For example, in Altman and Kamide's (1999) study, participants shift their gaze from an image of a boy to an image of a cake as soon as they hear "eat" in "The boy will eat the cake", as long as the cake is the only edible object in the scene (verb semantic constraint effect). This makes the visual world paradigm very useful for investigating the time-course of spoken language processing (see Huettig et al., 2011 for a review). What makes the visual world paradigm attractive for online eye tracking research is that the objects tend to be large and well-separated. This makes it easy to determine which object a participant is looking at even if accuracy is relatively low. Slim and Hartsuiker (2023) successfully replicated a visual world experiment with bilingual participants done by Dijkgraaf et al. (2017), although they reported a smaller effect size than that found in the original experiments and a delay of about 300 ms in the WebGazer results in relation to the spoken sentence presentation compared to the lab-based Dijkgraaf et al. (2017) study. Similarly, Prystauka et al. (2023) successfully replicated the verb semantic constraint effect first observed in the visual world paradigm by Altmann and Kamide (1999) and the lexical interference effect first observed by Kukona et al. (2014) where participants exhibit a small,

but consistent tendency to look at an object that is incompatible with the preceding verb they hear when a compatible colour word occurs (e.g., a black hat for the sentence “The grandfather will smoke the black pipe”). Prystauka et al. interpreted this as evidence that online eye tracking experiments are able to detect even relatively subtle effects, at least in the visual world paradigm.

In contrast, James et al. (2025) also attempted a replication of Altmann and Kamide’s (1999) results using WebGazer and failed to find evidence for the effect. This may be because their objects were relatively small and not always well separated (as in the original Altmann and Kamide study) as opposed to the four-quadrants design used by Prystauka et al. (2023) and highlights the accuracy issues that WebGazer has even in a design with relatively large stimuli. James et al. (2025) also tried to replicate the WebGazer experiment in the laboratory, using the built-in webcam of a Macbook laptop, and again failed to find evidence for the effect, suggesting that the issue is not due to the variability of participants’ devices and environments, but rather due to the inherent limitations of Webgaze. They did, however, succeed in replicating a number of other paradigms, including a visual world experiment reported by Ryskin et al. (2017), all using four-quadrant or two-halves designs. These limitations align with findings from a recent systematic evaluation of webcam-based eye-tracking algorithms, which highlighted that spatial drift, temporal latency, and browser-specific variability remain major obstacles to achieving laboratory-level precision (Thilderkvist & Dobslaw, 2024).

A number of other studies similarly find that online eye-tracking studies with designs analyzing very large portions of the screen as interest areas (as opposed to smaller, more

detailed areas of interest) tend to be successful when using WebGazer either on its own or integrated into an experiment platform such as Gorilla (Kandel & Snedeker, 2025; Slim et al., 2024; Steffan et al., 2024), even with young children or infants as participants.

Beyond these replication efforts, Bramlett and Wiener (2025) advanced the methodological foundations of web-based eye-tracking by introducing an open-science workflow for conducting and analysing online visual world experiments. Using WebGazer and Gorilla, they demonstrated that systematic calibration, preprocessing, and mixed-effects modelling can enable reproducible and interpretable analyses of online eye-tracking data, thereby bridging the gap between feasibility studies and standardized research practice. Furthermore, Patterson et al. (2025) outline best-practice guidelines for webcam-based eye tracking, including calibration routines, data-cleaning procedures, and transparent reporting.

Compared to the visual world paradigm, reading research requires higher precision, at least if the goal is to go beyond global measures such as the number of fixations on a passage. To put an error of 4.17º into perspective, one letter in 20 pt Courier New font, viewed at 60 cm distance on a 24-inch LCD monitor at a resolution of 1920x1080 pixels, subtends about .34º. This means that, with an error of 4.17º, the reported gaze location would be on average more than 12 letters away from the true location, making it very challenging to determine which word a participant is looking at and calculating accurate word-based fixation time measures. In comparison, the laboratory-based eye-tracker, with an error of .5 degrees, would on average report a gaze location 1.4 letters away from the true gaze location, which, in most cases, is still inside the word that the participant is fixating. As we describe below, this

problem can be mitigated using a larger font size, but this limits the length of the stimuli that can be fit on the screen.

If we wish to achieve an average accuracy of 1.4 letter spaces on an eye-tracker with an accuracy of 4.17º, each letter would have to subtend 3º of visual angle. If the visual angle subtended by a large screen is about 55º, we would only be able to fit 18 characters on a line, which is barely enough for three or four words. Additionally, the self-calibration mechanism used by WebGazer is not ideal for a reading paradigm, as reading does not require frequent mouse interaction – in fact, frequent use of the mouse may be indicative of a lack of focus on reading. To our knowledge, there has been only one attempt to collect data on eye movements during reading with WebGazer. The WebQAmGaze project (Ribeiro et al., 2024) involved the collection of 600 online eye movement recordings made using WebGazer while participants were reading paragraphs from existing lab-based eye movement corpora such as the MECO (Kuperman et al., 2025; Siegelman et al., 2022) in English, Spanish, German, and Turkish. The comparison of the WebGazer data with the original MECO data – which was obtained on high-precision eye trackers – showed some promising results, with clear correlations between the two data sets; however, the authors only present one analysis involving word properties (the word length effect), and, while the WebGazer data show a numerical tendency in the same direction as the MECO, the amount of noise in the data is substantial and may make it difficult to draw strong conclusions based on the WebGazer data alone. Additionally, the analysis is based on different sets of participants.

There have been several new online eye-tracking projects that have attempted to improve on the precision of WebGazer. Using customised software on a Google Pixel smartphone,

Valliapan et al. (2020) claim to have achieved an average error of about 0.6-1.0 degrees of visual angle, which would make their system's accuracy comparable to high-precision, lab-based eye trackers. While this is an exciting proof of concept showing what might be possible, the authors state that they are unable to release the software as it contains proprietary Google code, and, five years after the publication of the original article, no other study has used this software.

Taking a different approach, Saxena et al. (2023) used deep learning models to perform offline gaze detection on video recordings of participants' faces that were collected in an online experiment. They report an average error of about 2.4 degrees of visual angle, which is still a significant improvement compared to WebGazer. Saxena et al. 's software and data are all available as open source. However, the offline analysis approach raises the privacy concerns described above and, at the very least, requires specific consent from the participants for their faces to be recorded and the videos (which are personally identifiable information) to be stored on the researchers' systems in order to be analysed.

Another attempt at an improved software for online eye tracking comes from the Labvanced experimental platform (Finger et al., 2017). Similar to WebGazer, Labvanced performs all data processing on the participant's machine, but it has a more sophisticated head and gaze model based on machine learning. While the use of Labvanced requires a paid licence and the code is closed-access, the platform is currently publicly available and adding webcam eye-tracking to an existing experiment is very straightforward. To test the accuracy and precision of the Labvanced eye-tracking component, Kaduk et al. (2023) performed an experiment in which they recorded eye-tracking data simultaneously on an SR Research Eyelink 1000+ and

using a webcam with the Labvanced software while participants were performing a variety of mostly attention and eye-movement related tasks. They claim an average accuracy of 1.4º of error and a precision of 1.1º, both of which are very impressive compared to WebGazer, especially given that the Labvanced platform works on most devices with a webcam. Kaduk et al. included a number of tasks in their study but did not include a reading task.

From the publicly available software solutions for online eye-tracking reviewed above, Labvanced seems to be a promising one in terms of its reported accuracy and precision. Therefore, in the present study, we tested to what extent webcam eye-tracking can be used to study eye-movements during reading online. Similar to Kaduk et al. (2023), we directly compared the gaze position data obtained from Labvanced with the data from a high-precision SR Research Eyelink system using a co-registration setup. Additionally, similar to Angele et al. (2025), we attempted to replicate a key benchmark effect from the reading literature using webcams- the word frequency effect (Godwin et al., 2025). By simultaneously examining the lexical frequency effect with both a webcam and a standard lab eye-tracker, it was possible to test to what extent, if any, the two methods give comparable results.

We report three experiments: Experiment 1 was a lab-based co-registration study using English sentences presented as 3-4 lines of text on a large lab monitor. Each sentence contained a target word frequency manipulation and participants read the sentences while their eye movements were recorded by both an Eyelink 1000+ and the Labvanced system using an external webcam. An additional set of very short single-line sentences without any manipulation were included as a control. If the Labvanced system produces accurate and

precise eye-movement data, we should see a high correlation between the Eyelink and the webcam results. Additionally, if the Labvanced system is accurate and precise enough to detect evidence of the cognitive processing of word properties in word-based fixation time measures, we should observe effects of word frequency both in the Eyelink and the webcam data. In Experiment 2, we conducted a replication in Spanish and further optimised the presentation of sentences so that they appear only on two lines (at the top and the bottom of the screen). In this experiment, we again used a lab-based setup but with the stimuli presented on a laptop screen rather than a large computer monitor, which is more similar to the set-up that participants would use at home. Finally, in Experiment 3, we replicated Experiment 1 as a pure online experiment to test whether our findings would hold up in a more naturalistic experiment where the equipment and the environment was fully controlled by the participant.

1. Experiment 1

1.1. Method

1.1.1. Participants

Forty-two[1] English speakers from the London UCL area participated in return for a £10 payment (23 female). Participants' average age was 24.7 years (*SD*= 7.6 years; *range*: 18 –

[1] 8 more participants were tested, but their data was discarded for the following reasons: 3 due to poor calibration accuracy, 4 due to technical issues with the webcam recording, and 1 due to not being a fluent

49 years). Participants reported normal or corrected-to-normal vision (with contact lenses) and no prior diagnosis of reading disorders or any other neurological disorders. Participants wearing glasses were screened out due to incompatibility with the webcam recording. The study was approved by the Research Ethics Committee chair at the Department of Experimental Psychology, UCL (# EP_2024_01).

1.1.2. Materials

The reading materials consisted of 2 sets of stimuli: 60 single sentences with a lexical frequency manipulation (henceforth, “Frequency corpus”) and 40 single-line sentences (henceforth, “Single-line corpus”). The presentation order of the two corpora was counterbalanced across participants. Because a large font size had to be used to offset the inherent error of the webcam system, the frequency corpus sentences appeared on 3-4 lines of text across the whole screen. The single-line sentences all fitted on a single line of text and were used as a control to test how the webcam system performs in situations where tracking the y-position of the eye is not necessary.

1.1.2.1 Frequency-corpus. The stimuli were adapted from Vasilev et al.’s (2019) Experiment 1 (see Table 1). They consisted of 60 sentences that included a target word, which was manipulated to be of either high or low lexical frequency (measured with the SUBTLEX-UK corpus; van Heuven et al., 2014). An example sentence is given below (target word is underlined for illustration purposes but appeared normally in the text):

reader. One participant (#17) contributed data only to the Frequency corpus due to webcam recording issues in the second half of the study.

High-frequency condition:

Mrs. Clark is a social person who gets along with everybody.

Low-frequency condition:

Mrs. Clark is a chatty person who gets along with everybody.

The target words were an equal number of nouns and adjectives and differed significantly on lexical frequency, $t(110.81)= 29.33$, $p< 0.001$, but were matched on cloze predictability, $t(91.71)= -0.299$, $p = 0.765$, and word length ($p =1$). The sentences were on average 11.73 words (*SD*= 1.53) or 65.2 characters long (*SD*= 6.96). The position of the target word in the sentence differed for each item but was never the first or last word (Mean position= 6.21; *SD*= 1.65). The sentences were displayed on 3 to 4 lines of text; line breaks were created in such a way that the target word was never the first or last word on the line. Some of the original sentences were lightly edited (after the target word) to make them fit on the screen due to the bigger letters. Each sentence was followed by a Yes/No comprehension question (e.g., “Does Mrs. Clark get along with everybody?”). The presentation of sentences across the two conditions was counterbalanced with a Latin square design, such that participants saw either the high or low frequency version of each sentence, but across participants, both versions were presented equally often.

Table 1

Psycholinguistic Properties of the Target Words in Experiment 1 from the Frequency Corpus

	High-frequency target words				Low-frequency target words			
Variable	Mean	SD	Min	Max	Mean	SD	Min	Max
Lexical frequency (Zipf)	5.12	0.290	4.69	5.84	3.32	0.376	1.84	4
Cloze predictability	0.006	0.02	0	0.095	0.007	0.036	0	0.238
Word length	5.38	1.04	3	7	5.38	1.04	3	7

1.1.2.2 Single-line corpus. The stimuli consisted of 40 very short sentences that were specifically written to fit on a single line of text (thus avoiding the need to track the y-position of the eye). They were 3-5 words long (*Mean*=4.225; *SD*= 0.48 words) and contained 20.35 characters on average (*SD*= 0.863). Some examples are given below:

The man was fed up.

The cat saw the bird.

The green vase broke.

Each sentence was followed up by a Yes/No question (e.g., “Was the man pleased?”, “Did the cat see a dog?”).

1.1.3. Apparatus

Reference eye-movements were recorded with an Eyelink 1000+ at 1000 Hz (SR Research, Ontario, Canada). This was taken as the “ground truth” for comparing the accuracy of the webcam. However, as described above, the Eyelink1000+ system has an average spatial error of 0.25-0.5° itself, so the accuracy is not absolute. Viewing was binocular, but only the right eye was recorded. Participants' heads were stabilised with a chin-and-forehead rest. Data were recorded using a custom Python script (https://github.com/martin-vasilev/Eyelink_screen_recorder/). Calibration and validation of the Eyelink were performed using a 9-point grid at the start of the experiment. Participants were then recalibrated every 15 trials. The mean subject calibration error was 0.46° (*SD*= 0.10°; range: 0.27 - 0.65°).

Concurrent webcam eye-movements were recorded with a Logitech Streamcam at 60 Hz using the webcam interface at labvanced.com (Kaduk et al., 2023). Due to a technical issue, the first 3 participants were recorded at a 30 Hz sampling rate. The webcam interface did not achieve the full sampling rate, and the effective sampling rate was 56.35 Hz, on average (58.35 Hz if excluding the 3 participants recorded at 30 Hz). The webcam interface has a reported average accuracy of 1.45° (SD= 0.76°) (Kaduk et al., 2023). The calibration procedure consisted of 9 calibration screens- 7 screens consisting of dot calibration grids (approx. 12 targets each) and 2 consisting of smooth pursuit screens (see Kaduk et al., 2023

for more details). The points during calibration are separated into a training set and a validation set; a model is then built on the training set to calculate the subsequent accuracy on the validation set (Caspar Goeke, June 2025, personal communication). The average validation error reported by the system was 0.81° (*SD*= 0.76°) for X coordinates and 0.45° (*SD*= 0.36°) for Y coordinates. However, due to the different methods of calculation, we advise against comparing those values to the calibration accuracy of the Eyelink system.

To synchronise the two eye-movement data streams, UNIX timestamps were sent to the Eyelink computer (Kaduk et al., 2023), which gave the same time reference point. This was done at the start of the experiment and then every 5 trials to resynchronise the internal clocks. The Eyelink data was downsampled offline to the effective sampling rate of the webcam for each participant. This was done by identifying the Eyelink sample that occurred at the UNIX time point when the image was first acquired from the webcam. The two samples were then combined into the same dataset for analysis. Note that all results reported in this paper are based on the downsampled Eyelink data.

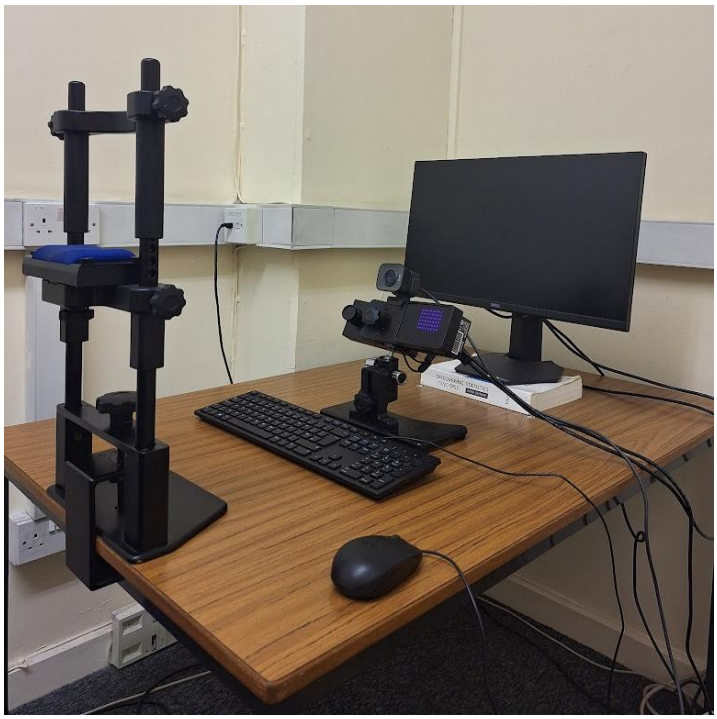

Figure 1. An illustration of the data acquisition setup in Experiment 1.

The experimental setup is illustrated in Figure 1. The stimuli were presented on a Windows 11 PC with a 24-inch LCD Monitor (resolution: 1920 x 1080 pixels; refresh rate: 165Hz; physical screen size: 53 cm x 30 cm). The distance between the eye and the Eyelink 1000+ was 55 cm, and the distance between the eye and the webcam was 50 cm. The Logitech Streamcam was mounted on top of the Eyelink, such that it did not obstruct the camera, illuminator module, or block the participant's view of the monitor. The eye-to-monitor distance was 82 cm. At this distance, each pixel subtended 0.0187° horizontally and 0.0192° vertically.

The sentences were formatted in a 60pt. monospaced font that appeared as black text on a white background. The text was offset at 125 px horizontally and 150 px vertically. The text was double-spaced and aligned to the left. The width of each letter was 78 px (1.46°), matching the reported spatial accuracy of the webcam (1.45°; Kaduk et al., 2023).

1.1.4. Procedure

Participants were tested in a small windowless room illuminated by an overhead lamp (which was not in view of either camera). Participants provided informed written consent and were positioned on a chin and forehead rest (see Figure 1). Once the webcam detected the position of the participants' face, the chinrest was removed, and participants completed the webcam calibration. The chinrest was removed during this procedure because participants needed to shift their head position slightly during the calibration process (see Kaduk et al., 2023). Afterwards, participants were repositioned on the chinrest and were calibrated on the Eyelink eye-tracker[2]. Then, the experiment commenced.

The study started with 6 practice trials, followed by the 100 experimental trials (60 in the Frequency corpus, 40 in the Single-line corpus). The sentences in the two corpora were blocked, but within each block, the sentences were presented in a pseudo-random order. Each trial began with a fixation cross (82 x 82 pixels) placed at the first letter of the sentence. After the webcam detected the participant's gaze on the area around the fixation cross, the sentence was presented on the screen. The experimenter manually presented the sentence with a

[2] It should be noted that, as far as we can tell, Kaduk et al. (2023) did not use a chinrest for recording. We opted to do so to minimise head movements and increase recording accuracy.

keypress if the gaze-contingent presentation of the sentence was not triggered within a few seconds. Participants then read the sentence silently at their own pace and clicked the left button of the mouse to indicate they had finished reading. After each sentence, there was a True/ False comprehension question, which participants answered with a mouse click. A drift check was completed every 5th trial. For the Eyelink, this was a single target presented at the centre of the screen; for the webcam interface, this was a 7-point grid (Kaduk et al., 2023). The Eyelink was recalibrated every 15 trials. The study took about 45 minutes to complete.

1.1.5. Data Analysis

The raw data was first cleaned by removing samples occurring outside the screen area (2.09%), as well as any samples that occurred during a blink (5.09%). For the Eyelink, blinks were defined as any samples where the pupil size was 0. For the webcam interface, a confidence rating (0-1) indicated whether the eyelid was open. We selected only cases where the rating was >0.

To assess eye-tracking consistency between the two signals, we computed Pearson's correlation coefficient between the webcam and Eyelink gaze coordinates separately for the horizontal (x) and vertical (y) dimensions. This approach was inspired by Kaduk et al. (2023), who reported high overall correlations between the two signals. Furthermore, to estimate the relative error across screen regions, we divided the display into a 3 × 3 grid based on the monitor resolution. Each gaze sample was assigned to one of nine equally sized sections (Section 1-9 from top-left to bottom-right) and the difference between the webcam and Eyelink gaze locations was calculated for each region.

To analyse reading behaviour, the raw eye samples from both systems were parsed into fixations using Engbert and Kliegl's (2003) algorithm, as implemented in the "saccades" R package (von der Malsburg, 2019). We used the default λ= 6 parameter for saccade detection sensitivity and a 3-sample moving window smoothing of gaze coordinates to reduce noise. This algorithm has previously been shown to be sensitive in low-sampling resolution applications for detecting the lexical frequency effect (Angele et al., 2025). Fixations shorter than 80 ms that were within 1 character distance of a temporally adjacent fixation were merged into that adjacent fixation (Frequency corpus: 0.38% [webcam], 1.76% [Eyelink]; Single-line corpus: 0.56% [webcam], 2.32% [Eyelink]); any remaining fixations under 80 ms were discarded as outliers (Frequency corpus: 3.46% [webcam], 16.17% [Eyelink]; Single-line corpus: 3.94% [webcam], 15.82% [Eyelink])[3]. Additionally, fixations longer than 1000 ms were also discarded as outliers (Frequency corpus: 2.93% [webcam], 0.47% [Eyelink]; Single-line corpus: 4.22% [webcam], 0.52% [Eyelink]).

From the raw fixation data, we computed the following reading measures: *first fixation duration* (FFD; the duration of the first fixation on a word during first-pass reading); *single fixation duration* (SFD; the duration of a fixation on a word when it is fixated exactly once during first-pass reading), *gaze duration* (GD; the sum of all fixation durations on a word during first-pass reading), and *total viewing time* (TVT; the sum of all fixations on a word, including when a word is re-fixated during second-pass reading). We trimmed durations

[3] There were more fixations < 80ms detected for the Eyelink data compared to the webcam data. Because the webcam data is noisier, there are fewer small saccades that are detected by the algorithm. As a result, the data tends to be grouped into fewer, but longer fixations. The Eyelink data, on the other hand, has more spatially consistent signal, which makes the detection of more of these smaller saccades possible. While this can be addressed by increasing the detection threshold (λ) for the Eyelink data, we opted not to do so in the spirit of treating both signals in the same way.

above 2000 ms for GD and 3000 ms for TVT as outliers (Frequency corpus: 0.22% [webcam], 0.09 % [Eyelink]; Single-line corpus: 4.02% [webcam], 0.59% [Eyelink]).

We analysed the data for the frequency corpus by fitting Bayesian linear mixed models using the brms package (Bürkner, 2017, 2018, 2021) in R[4]. These models included the word frequency condition (low, represented as -.5, vs. high, represented as .5) as well as the eye-tracker used (with the Eyelink represented as -.5 and the Webcam represented as .5). We included the maximum random effects structure (i.e., random intercepts, random slopes for Frequency and Tracker, and their interaction over both subjects and items, Barr et al., 2013). We used the Gaussian distribution to model log-transformed fixation times with the default priors suggested by brms for the fixed effects and random effects, except for the slope coefficients were given weakly informative priors of *b ~Normal(0, 1)* to rule out implausibly large effects.

Each model was fitted using four chains with 5000 iterations each with 1000 warmup iterations. The models converged successfully (all R-hats = 1.00). We report the estimates (*b*) and the 95% Bayesian Credible Intervals (95% CrIs) based on the posterior distribution of

[4] The versions of R and all packages used are as follows: R (Version 4.5.2; R Core Team, 2025) and the R-packages *conflicted* (Version 1.2.0; Wickham, 2023), *dplyr* (Version 1.1.4; Wickham et al., 2023), *emmeans* (Version 2.0.0; Lenth & Piaskowski, 2025), *forcats* (Version 1.0.1; Wickham, 2025), *ggplot2* (Version 4.0.0; Wickham, 2016), *gt* (Version 1.1.0; Iannone et al., 2025), *here* (Version 1.0.2; Müller, 2025), *knitr* (Version 1.50; Xie, 2015), *lubridate* (Version 1.9.4; Grolemund & Wickham, 2011), *modelsummary* (Version 2.5.0; Arel-Bundock, 2022), *papaja* (Version 0.1.4; Aust & Barth, 2022), *purrr* (Version 1.2.0; Wickham & Henry, 2025), *qs2* (Version 0.1.5; Ching, 2025), *readr* (Version 2.1.5; Wickham et al., 2024), *sjPlot* (Version 2.9.0; Lüdecke, 2025), *stringr* (Version 1.6.0; Wickham, 2025b), *tibble* (Version 3.3.0; Müller & Wickham, 2025), *tidybayes* (Version 3.0.7; Kay, 2024), *tidyr* (Version 1.3.1; Wickham, Vaughan, et al., 2024), *tidyverse* (Version 2.0.0; Wickham et al., 2019), and *tinylabels* (Version 0.2.5; Barth, 2025).

each model parameter and indicate whether 0 is excluded from the 95% CrI. The exclusion of 0 from the 95% CrI was the main method of inference used to determine the presence of an effect. Additionally, we report the proportion of the posterior on the same side of 0 (P(b>0) or P(b<0)) and its ratio to the opposite-side proportion as a directional evidence ratio ($ER_{b>0}$ or $ER_{b<0}$), which can be interpreted as a Bayes factor for the directional hypothesis. Finally, we report Bayes factors (Jeffreys, 1961) for the point-null hypothesis (b = 0), calculated using the Savage–Dickey density ratio method (Dickey & Lientz, 1970; Morey et al., 2011). It is important to note that each of these statistics is a valid way to summarize the evidence provided by the data, but, for simplicity, we will focus on the 95% CI and whether it excludes 0 in our verbal description of the results.

1.2. Results and Discussion

Participants' mean comprehension accuracy was 96.3% in the Frequency corpus (*SD*=19.0%; *range*= 73.3-100%) and 97.1% in the Single-line corpus (*SD*= 16.9%; *range*= 85 - 100%). This shows that participants were reading for comprehension. Average sentence reading times were 3708 ms for the Frequency corpus (SD= 2026 ms) and 1799 ms for the Single-line corpus (SD= 1043 ms).

1.2.1. Comparison of raw sample positions

To measure the relative consistency in the two signals, the two eye-tracker samples were correlated separately for the x and y dimensions of the screen (see Figure 2). The average correlations were r= 0.78 (x-dimension) and r= 0.76 (y-dimension) for the frequency corpus.

These values align closely with the *r*= 0.83 (X) and *r*= 0.84 (Y) reported by Kaduk et al. (2023) in a free-viewing task. The average correlations were *r*= 0.86 (x-dimension) and *r* =0.23 (y-dimension) for the single-line corpus. The low correlation for the Y-axis likely shows that the webcam is less able to capture small deviations in the y position of the eye in a task where vertical eye movements are not needed.

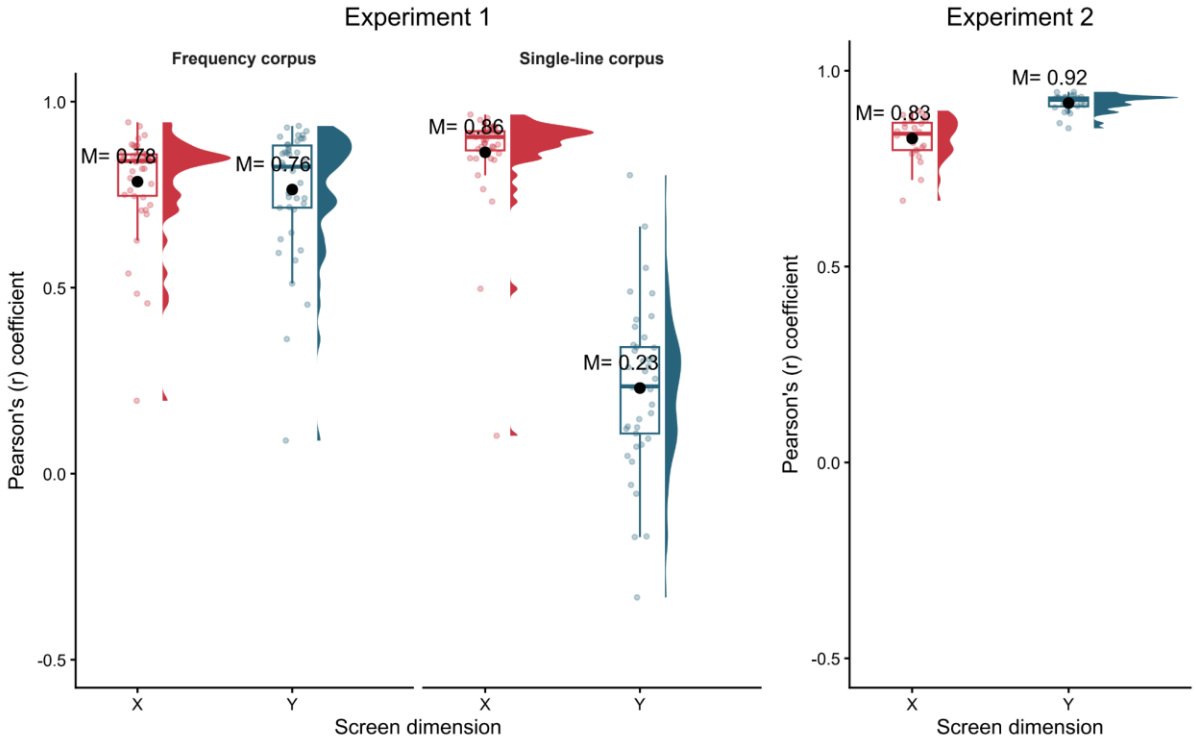

Figure 2. Pearson's *r* correlations between the Eyelink and webcam gaze sample position in Experiments 1 and 2, calculated separately for the x- and y-dimension. Each dot represents the average correlation for a participant in the study. The black dot represents the average across all participants.

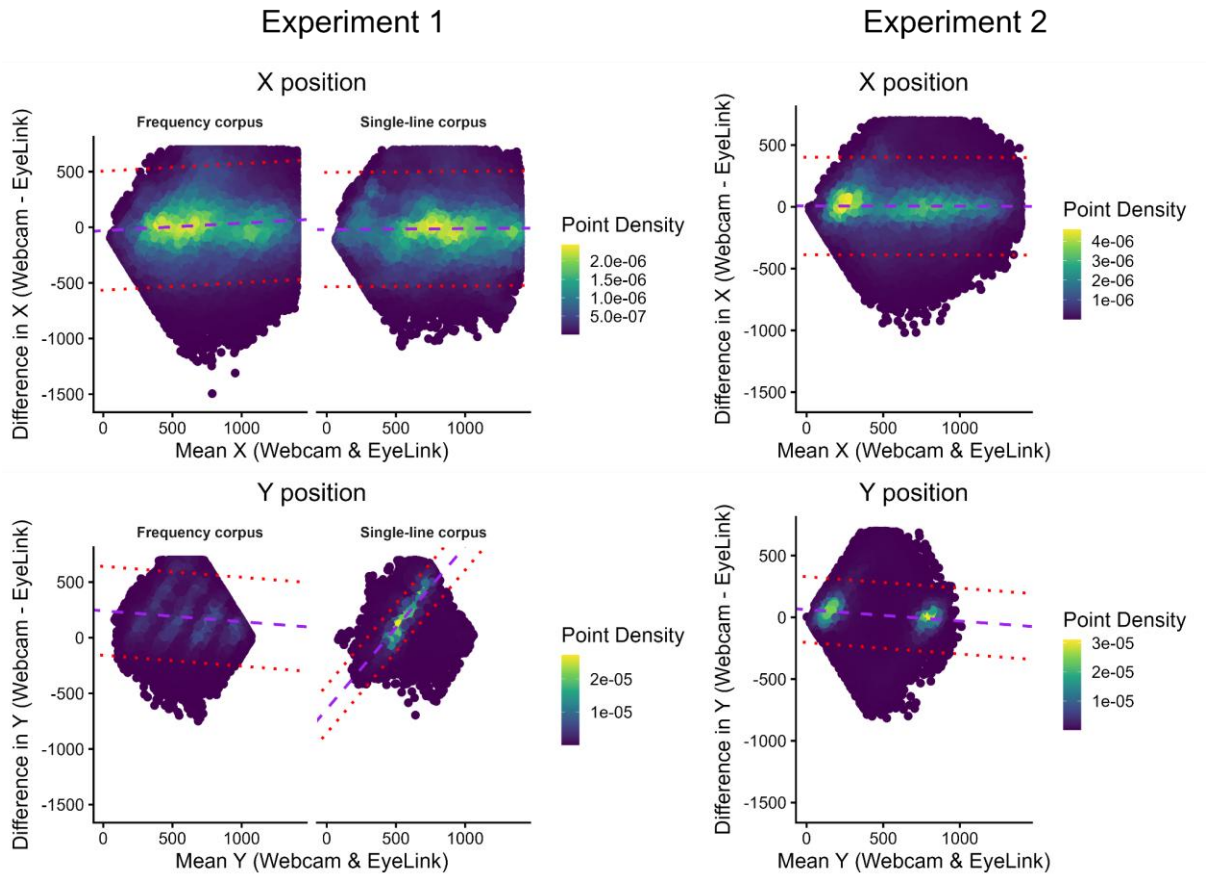

Figure 3. Bland and Altman (1999) plots showing the difference between the webcam and Eylink raw eye positions for the X-axis (top row) and Y-axis (bottom row), plotted against the mean value for both cameras. Dotted lines indicate the linear regression slope and 95% CI limits. Data from Experiment 1 is shown on the left and data from Experiment 2 is shown on the right.

The agreement between the two systems was further explored with Bland and Altman (1999) plots. These are often used for comparing a new measurement system against an existing

"gold" standard. The difference in raw eye positions between the two systems is plotted against their mean value. If the new system shows no error or bias, the differences will be distributed around 0 and exhibit a horizontal slope, with 95% of the values falling within the confidence limits. If the slope of the scatterplot is not 0, this shows a "proportional bias"- i.e., the difference between two systems changes with the magnitude of measurement (Ludbrook, 2010).

As Figure 3 shows, the webcam and Eyelink agreed well on estimating the X position of the eye, showing little evidence of systematic bias. However, the confidence limits were very wide, indicating considerable uncertainty in the measurement. On the other hand, the Y position of the eye was consistently overestimated on the Frequency corpus. For the Single-line corpus, the opposite pattern was found- the webcam consistently underestimated low Y positions and overestimated high Y positions. As the Single-line corpus did not require participants to move their eyes vertically, this suggests that the webcam does not differentiate well between small fluctuations in the Y position of the eye.

The average magnitude of the webcam error relative to the Eyelink was 0.25° (*SD*= 5.08°) for the x-dimension and 3.50° (*SD*= 3.83°) for the y-dimension (see Figure 4a). This shows that the webcam estimated the average x position with little bias, though there was considerable variance around the mean. The y position was consistently overestimated by the webcam; there was a similar (though smaller) spread of values around the mean. We also compared the magnitude of the error based on the screen area where the samples occurred (separated in 9 rectangles in a 3 x 3 grid). As Figure 5a shows, the x position was recorded with the least

amount of bias at the centre of the screen. On the other hand, the y position was estimated with the least amount of bias at the bottom of the screen.

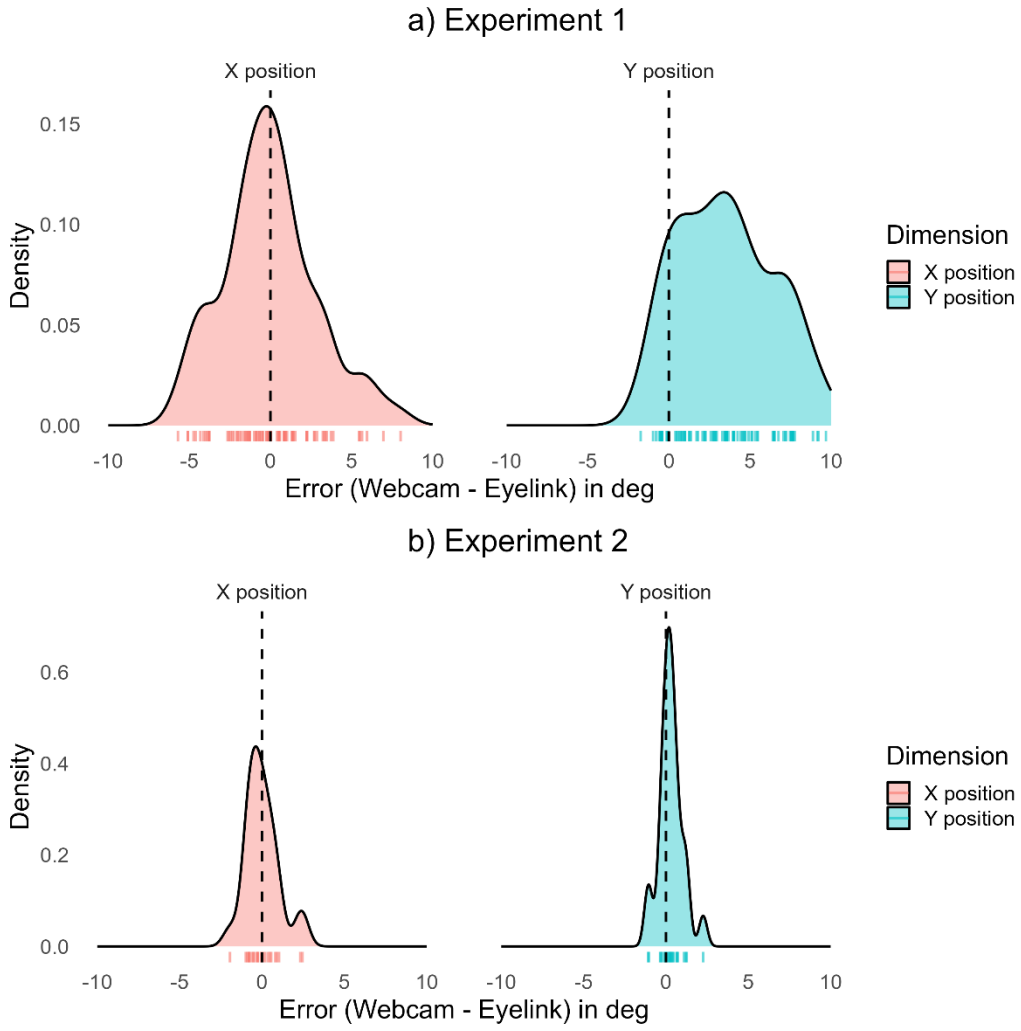

Figure 4. Average positional difference between the webcam and Eyelink recordings in Experiment 1 (**a**) and Experiment 2 (**b**), expressed in degrees of visual angle. The data was aggregated over subjects before plotting. Negative values indicate an underestimation of the Eyelink position by the webcam. A vertical dotted line at 0 is plotted to aid interpretation.

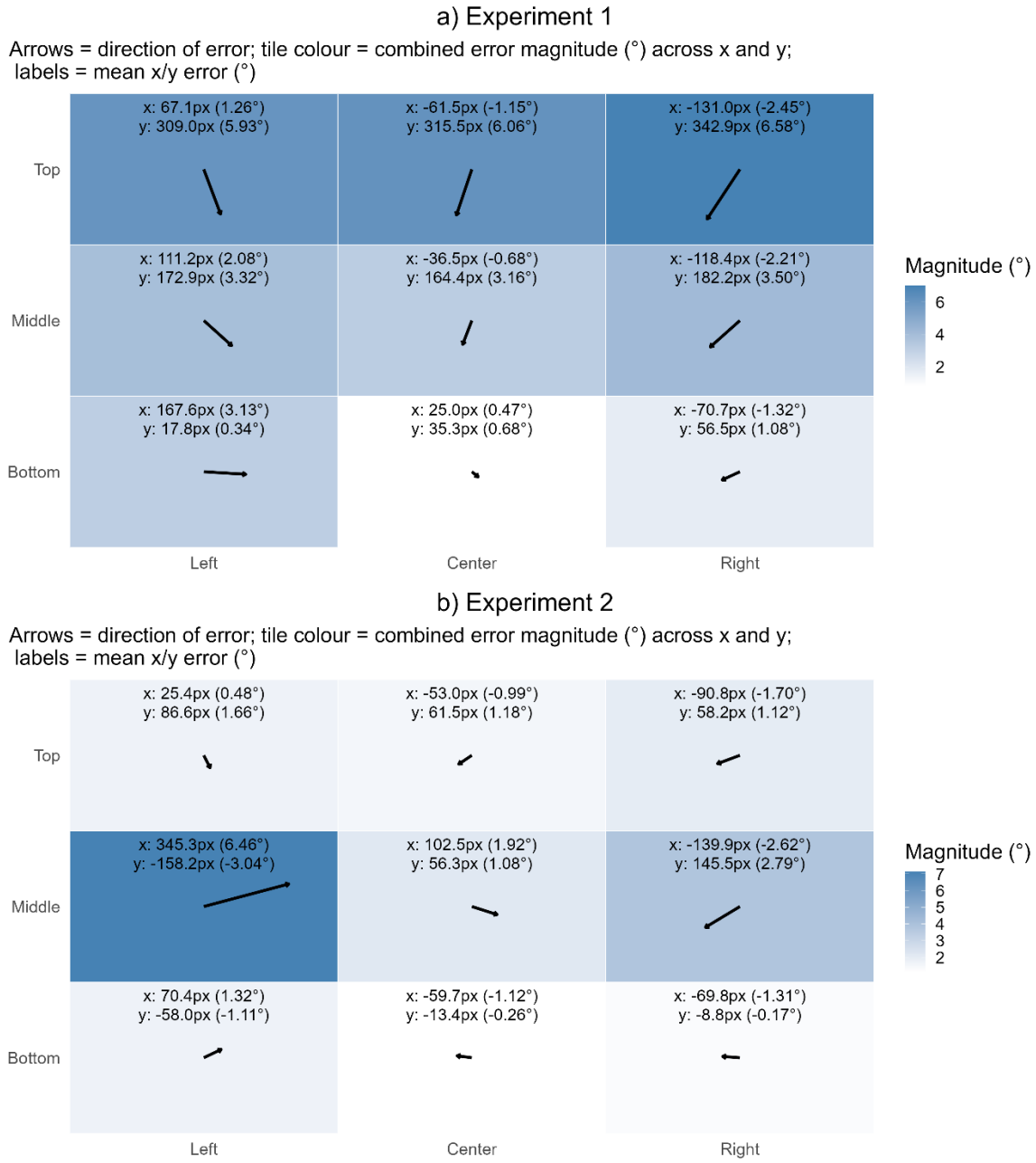

Figure 5. Mean directional error per screen section in Experiment 1 (**a**) and Experiment 2 (**b**). Error was calculated by dividing the screen area into a 3 x 3 grid and calculating the positional error between the webcam and Eyelink (arrows indicate positional bias).

1.2.2. Analysis of the target word lexical frequency effect

We now turn to analysing fixation durations on the target word to check for evidence of the lexical frequency effect. Table 2 shows the means and standard deviations for the four fixation time measures on target words in Experiment 1, while Table 3 shows the corresponding LMM model estimates.

Overall, low-frequency words elicited longer fixation times than high-frequency words across both devices, with the 95% CrI excluding 0 (FFD: $b = 0.06$, 95% CrI $[0.02,\ 0.10]$, $P(b > 0) = 1.00$, $ER_{b>0} = 379.95$, $BF_{10} = 0.88$; SFD: $b = 0.06$, 95% CrI $[0.02,\ 0.10]$, $P(b > 0) = 1.00$, $ER_{b>0} = 306.69$, $BF_{10} = 0.88$; GD: $b = 0.09$, 95% CrI $[0.04,\ 0.13]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} = 17.87$; TVT: $b = 0.11$, 95% CrI $[0.06,\ 0.17]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} = 42.70$). Fixation times were also longer with the Webcam than with the Eyelink (FFD: $b = 0.41$, 95% CrI $[0.33,\ 0.49]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; SFD: $b = 0.41$, 95% CrI $[0.32,\ 0.49]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; GD: $b = 0.43$, 95% CrI $[0.35,\ 0.52]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; TVT: $b = 0.56$, 95% CrI $[0.45,\ 0.67]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$). This may be due to the higher spatial resolution of the Eyelink system which enabled the algorithm to detect shorter fixations that may have been missed in the Webcam data due to its greater variance in gaze positions. Importantly, the Frequency × Tracker interaction was not credibly different from 0 (FFD: $b = -0.04$, 95% CrI $[-0.11,\ 0.03]$, $P(b < 0) = 0.85$, $ER_{b<0} = 5.73$, $BF_{10} = 0.06$; SFD: $b = -0.03$, 95% CrI $[-0.11,\ 0.05]$, $P(b < 0) = 0.78$, $ER_{b<0} = 3.50$, $BF_{10} = 0.05$; GD:

$b = -0.08$, 95% CrI $[-0.17,\ 0.007]$, $P(b < 0) = 0.96$, $ER_{b<0} = 27.02$, $BF_{10} = 0.24$; TVT: $b = -0.06$, 95% CrI $[-0.16,\ 0.03]$, $P(b < 0) = 0.91$, $ER_{b<0} = 10.31$, $BF_{10} = 0.12$), indicating that the frequency effect was comparable across devices.

The absence of an interaction between frequency and tracker is not in itself positive evidence for a frequency effect in the Webcam data. We calculated posterior distributions for the estimated marginal means of the difference between low and high frequency target words and found that the 95% credible intervals for these differences in the Webcam data all included 0, while the credible intervals for the Eyelink data all excluded 0 in FFD (Webcam: *M* = 0.04 95% *CrI* = [-0.02, 0.10]; Eyelink: *M* = 0.08 95% *CrI* = [0.04, 0.12]), SFD (Webcam: *M* = 0.04 95% *CrI* = [-0.03, 0.11]; Eyelink: *M* = 0.07 95% *CrI* = [0.03, 0.12]), GD (Webcam: *M* = 0.04 95% *CrI* = [-0.03, 0.12]; Eyelink: *M* = 0.13 95% *CrI* = [0.08, 0.17]), and TVT (Webcam: *M* = 0.08 95% *CrI* = [-0.00, 0.16]; Eyelink: *M* = 0.14 95% *CrI* = [0.08, 0.20]), although the CrI for TVT included 0 very narrowly for the webcam data. This indicates that, while the webcam data is consistent with a frequency effect of a similar magnitude to that observed with the Eyelink, the evidence is not sufficiently strong to confirm the presence of frequency effect based on the Webcam data alone using the criterion of the 95% credible interval excluding 0. We assume that this is due to the higher amount of noise in the Webcam data.

Table 2

Mean first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on target words in Experiment 1, grouped by eye-tracker and target word frequency condition.

		FFD		SFD		GD		TVT	
Tracker	Frequency	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Webcam	low	281	173	287	174	352	239	559	390
Webcam	high	265	159	272	167	333	223	522	399
Eyelink	low	190	133	197	146	236	163	342	253
Eyelink	high	176	124	183	135	207	142	299	230

Note: All means and standard deviations are reported in milliseconds (ms).

Table 3

Coefficient estimates and 95% credible intervals (in square brackets) from Bayesian linear mixed models fitted to first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on target words in Experiment 1.

	FFD	SFD	GD	TVT
Intercept	**5.298**	**5.317**	**5.472**	**5.821**
	[**5.200, 5.397**]	[**5.208, 5.421**]	[**5.368, 5.576**]	[**5.707, 5.931**]
Frequency	**0.058**	**0.057**	**0.085**	**0.112**
	[**0.017, 0.097**]	[**0.013, 0.098**]	[**0.040, 0.130**]	[**0.056, 0.167**]
Tracker	**0.413**	**0.405**	**0.433**	**0.553**
	[**0.334, 0.492**]	[**0.319, 0.491**]	[**0.344, 0.524**]	[**0.443, 0.668**]
Frequency × Tracker	-0.037	-0.031	-0.080	-0.064
	[-0.109, 0.034]	[-0.108, 0.048]	[-0.170, 0.010]	[-0.159, 0.030]

Note: Effects for which the 95% CrI excludes 0 are highlighted in bold.

Table 4

Coefficient estimates and 95% credible intervals (in square brackets) from Bayesian linear mixed models fitted to first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on all words (except for the first and last in each sentence) in Experiment 1.

	FFD	SFD	GD	TVT
Intercept	**5.267**	**5.278**	**5.401**	**5.647**
	[**5.182, 5.355**]	[**5.189, 5.363**]	[**5.314, 5.482**]	[**5.563, 5.728**]
Zipf frequency	**-0.037**	**-0.043**	**-0.082**	**-0.119**
	[**-0.045, -0.028**]	[**-0.053, -0.033**]	[**-0.093, -0.070**]	[**-0.134, -0.103**]
Tracker	**0.402**	**0.396**	**0.460**	**0.548**
	[**0.343, 0.461**]	[**0.332, 0.460**]	[**0.386, 0.536**]	[**0.474, 0.623**]
Frequency × Tracker	**0.019**	**0.022**	**0.013**	0.013
	[**0.009, 0.029**]	[**0.012, 0.033**]	[**0.002, 0.025**]	[-0.000, 0.026]

Note: Effects for which the 95% CrI excludes 0 are highlighted in bold.

1.2.3. Analysis of corpus lexical frequency

By including all words in every sentence into the analysis, rather than only the target words, we can increase the number of observations. Of course, this comes at the cost of increased variability due to differences in word length, predictability, and other lexical and contextual

factors, but, overall, this approach is likely to have a better chance to detect the word frequency effect. Table 4 shows the results of these "corpus-style" Bayesian linear mixed model analyses with Zipf frequency as a continuous predictor, tracker as a two-level factor (as in the target-word analysis) and the interaction between the two predictors for the four fixation time measures on all words in the sentences (except for the first and last word in each sentence) in Experiment 1. Note that this analysis contains the combined items from both the Frequency corpus and the Single-line corpus (100 items in total).

As expected, we observed a frequency effect in that a higher Zipf frequency (i.e., more frequent words) was associated with shorter fixation times across all measures (FFD: $b = -0.04$, 95% CrI $[-0.05,\ -0.03]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$; SFD: $b = -0.04$, 95% CrI $[-0.05,\ -0.03]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$; GD: $b = -0.08$, 95% CrI $[-0.09,\ -0.07]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$; TVT: $b = -0.12$, 95% CrI $[-0.13,\ -0.10]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$). We again observed longer mean fixation times with the Webcam than with the Eyelink (FFD: $b = 0.40$, 95% CrI $[0.34,\ 0.46]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; SFD: $b = 0.40$, 95% CrI $[0.33,\ 0.46]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; GD: $b = 0.46$, 95% CrI $[0.39,\ 0.54]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; TVT: $b = 0.55$, 95% CrI $[0.47,\ 0.62]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$).

There was a small, but credible Frequency × Tracker interaction in FFD ($b = 0.02$, 95% CrI $[0.0093,\ 0.03]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} = 2.19$), SFD ($b = 0.02$, 95% CrI $[0.01,\ 0.03]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} = 14.70$), and GD ($b =$

0.01, 95% CrI $[0.0021,\ 0.02]$, $P(b > 0) = 0.99$, $ER_{b>0} = 95.39$, $BF_{10} = 0.09$), but not in TVT ($b = 0.01$, 95% CrI $[-0.0001,\ 0.03]$, $P(b > 0) = 0.97$, $ER_{b>0} = 37.93$, $BF_{10} = 0.05$). This interaction indicates that the frequency effect was stronger in the Eyelink data compared to the Webcam data. To show the marginal effect of zipf frequency within each device directly, we summarized the slope of the frequency effect using `emtrends` from the `emmeans` package. We found consistent evidence for the frequency effect in all measures for both the Eyelink and the Webcam (FFD: Eyelink: *M* = -0.05 95% *CI* = [-0.06, -0.04]; Webcam: *M* = -0.03 95% *CI* = [-0.04, -0.02], SFD: Eyelink: *M* = -0.05 95% *CI* = [-0.06, -0.04]; Webcam: *M* = -0.03 95% *CI* = [-0.04, -0.02], GD: Eyelink: *M* = -0.09 95% *CI* = [-0.10, -0.08]; Webcam: *M* = -0.07 95% *CI* = [-0.09, -0.06], TVT: Eyelink: *M* = -0.12 95% *CI* = [-0.14, -0.11]; Webcam: *M* = -0.11 95% *CI* = [-0.13, -0.09]). Overall, this demonstrates that, when there is sufficient data, the corpus frequency effect can be detected even in the noisier Webcam data.

1.2.4. Summary and conclusions

The results of Experiment 1 can be summarised as follows. First, the time series correlations of the webcam and Eyelink data showed a strong association, replicating Kaduk et al.'s (2023) results in a free-viewing task. Second, the webcam gaze coordinates generally showed much larger variation compared to the Eyelink gaze coordinates. While little systematic bias was found in estimating x positions, y positions were generally overestimated for the Frequency corpus (multi-line reading). Third, reliable lexical frequency effects on the target word were found only for the Eyelink data. The effect in the webcam data was numerically similar, though the 95% CrIs all included 0, suggesting that the effect is not

statistically reliable. To boost statistical power, we conducted a corpus frequency analysis of all words in the sentence (instead of just the target word). This analysis revealed reliable corpus frequency effects in the webcam data. This suggests that corpus frequency effects can be detected with a larger number of observations.

Overall, these results show some promise in using webcam eye-tracking for reading research. However, they also demonstrate some challenges, such as the considerable additional noise in estimating gaze coordinates and systematic errors in tracking y-positions. If webcam eye-tracking is to be practically useful, some form of multiline reading is necessary to ensure that the stimuli can fit on the screen, given that letters need to be much bigger to counter the inherent error in webcam tracking. The current study presented the Frequency corpus items over 3-4 lines of text; however, using two lines of text may reduce the vertical error associated with mapping fixations to the text. In Experiment 2, we conducted a conceptual replication of the study in Spanish, using sentences with a lexical frequency manipulation spanning 2 lines of text. The study was also conducted on a laptop (rather than a lab monitor), which is closer to the setup that participants would have at home.

2. Experiment 2

2.1 Method

2.1.1. Participants

Twenty-four native Spanish speakers (15 female) from Universidad Nebrija participated in the study in exchange for a 6€ payment for a 30-minute experiment. Their average age was 26.4 years (SD = 3.4; range: 21–31 years). All participants reported normal or corrected-to-normal vision, including those wearing contact lenses, and had no history of reading or neurological disorders. Participants who required glasses to read were excluded from the study. The study was approved by the Ethics Committee at Universidad Nebrija (Ref. UNNE-2023-0031).

2.1.2. Materials

The reading materials consisted of 200 sentences, arranged in 100 pairs. Each pair shared the same sentence structure at the beginning but differed in the lexical frequency of a single target word (high or low frequency) and sometimes in the context after the target word, based on the EsPal database (Duchon et al., 2013) for Spanish (see Table 5 for more information). The sentences were written specifically for the experiment and had a maximum length of 50 characters, including spaces and punctuation. The target word, always a noun, appeared in the middle of the sentence. An example item is given below:

High-frequency condition:

Se podía ver que el país iba a mejor.

Low-frequency condition:

Se podía ver que el humo se había disipado.

Target words ranged from 4 to 9 letters in length and were carefully balanced across grammatical properties: half were feminine nouns and half masculine nouns, and within each gender category, half were singular and half were plural. Thus, each frequency condition (high vs. low) contained a balanced number of targets across gender and number.

The stimuli were divided into two lists: List A and List B. Each list contained 100 sentences: 50 with high-frequency target words and 50 with low-frequency target words. The other list contained the complementary items of each pair (i.e., the opposite frequency condition). Participants with odd-numbered IDs were assigned to List A, and even-numbered participants to List B, ensuring that each participant saw only one version of each sentence pair. Within each list, sentence presentation order was randomized for each participant.

Sentences were displayed across two lines (no more than 25 characters per line) using a monospaced Courier New font, size 48, in black text on a white background. One-third of the sentences were followed by a Yes/No comprehension question. Additionally, participants completed a short practice session before starting the experiment, consisting of five example sentences and three comprehension questions.

Table 5

Psycholinguistic Properties of the Target Words in Experiment 2 from the database "EsPal" (Duchon et al., 2013)

	High-frequency target words				Low-frequency target words			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Lexical frequency (Zipf)	5.21	0.29	4.70	5.87	3.94	0.04	3.80	4
Word length	6.5	1.51	4	9	6.5	1.51	4	9

2.1.3. Apparatus

Eye movements were recorded simultaneously using an Eyelink Portable Duo eye tracker (SR Research, Ontario, Canada) at 1000 Hz, and a Logitech Streamcam webcam at 60 Hz via the Labvanced platform (Kaduk et al., 2023) (see Figure 6 for an illustration). The webcam achieved an effective sampling rate of 57.24 Hz on average (*SD*= 1.82 Hz; *range*: 53.24-59.72 Hz). Participants were seated in front of the laptop and used a chinrest to stabilize head movement. The eye-tracking setup was run on a Windows 11 MSI Raider RGB laptop with a 24-inch LCD screen (resolution: 1920 × 1080 pixels; refresh rate: 144 Hz; physical size: 35 x 20 cm). The eye-to-monitor distance was 63 cm. The webcam was positioned 58 cm from the

eye and the Eyelink camera was positioned 53 cm from the eye. In this setup, each pixel subtended 0.0162° horizontally and 0.0167° vertically. Each letter had a width of 55 pixels (0.89°).

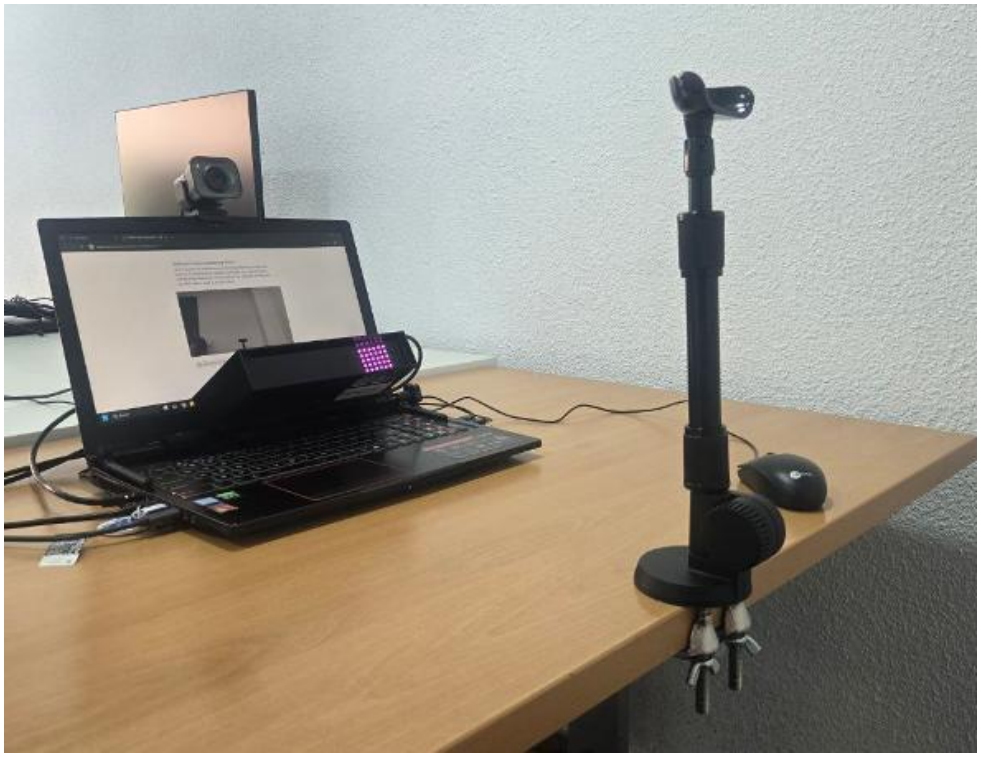

Figure 6. Apparatus set-up in Experiment 2.

The calibration procedure included two stages: first, a Labvanced webcam calibration with head movement tracking across various angles, followed by a 9-point calibration and validation procedure with the Eyelink system while the participant remained in the chinrest. The signals from the Eyelink and the webcam were synchronized offline using UNIX timestamps at the start of the experiment, and the Eyelink data were downsampled offline to match the effective webcam sampling rate, following the same procedure as Experiment 1.

2.1.4. Procedure

Participants gave written informed consent prior to participation. They were tested in a quiet room. After completing the dual calibration procedures (webcam and Eyelink), the experiment began. Participants first completed five practice trials, followed by the 100 experimental trials.

Each trial began with a fixation cross (Shape ABC from Experiment 2 in Thaler et al., 2013, with a black outer circle, white cross, and red inner circle) at the position of the first letter of the sentence. After 250 ms, the sentence appeared. The sentences were presented in two left-aligned horizontal lines, each with a maximum length of 725 units (31.72 cm), such that each line was offset by 25 units (1.09 cm) from the left and right edges of the screen, respectively, occupying 90.63% of the total screen width in an 800 × 450 unit space. The first line was presented at the top of the screen, with a y-offset of 25 units (1.11 cm; 5.6% of screen height) from the top of the screen, and the second line was presented at a y-offset of 360 units (16.00 cm; ~80% of screen height) from the top of the screen. Participants read the sentence silently and clicked anywhere on the screen to continue to the next trial. In one-third of the trials, a comprehension question followed the sentence, requiring a Yes/No mouse response. The entire session lasted between 20 and 30 minutes, including approximately 5-10 minutes for calibration and 10-15 minutes for the reading task.

2.1.4. Data analysis

The data was analysed in the same way as Experiment 1. Samples that contained blinks (2.44%) or were otherwise outside the screen area (1.56%) were excluded from the data. Fixations shorter than 80 ms that were within 1 character distance of a temporally adjacent fixation were merged into that adjacent fixation (0.21% [webcam], 1.65% [Eyelink]); any remaining fixations under 80 ms were discarded as outliers (4.30% [webcam], 16.2% [Eyelink]). Additionally, fixations longer than 1000 ms were also discarded as outliers (1.23% [webcam], 0.31% [Eyelink]).

2.2. Results and Discussion

2.2.1. Comparison of raw sample positions

Participants' average comprehension accuracy was 96.1% (*SD*= 19.35%; range: 83.3%-100%), indicating that they read the sentences for understanding. The average trial duration was 2143 ms (*SD*= 899 ms). Person's *r* correlations between the Eyelink and the webcam data are shown in Figure 2. The sample correlations were *r*= 0.83 for the x-axis and *r*= 0.92 for the y-axis, indicating a slightly stronger association compared to the Experiment 1 results, particularly for the y-dimension.

The Bland and Altman (1999) plots (see Figure 3) showed very similar results to Experiment 1: there was little systematic bias in estimating the x-dimension, though the webcam again showed considerable variance in gaze positions. The y-dimension showed a weak negative trend, indicating that y-coordinates at the top of the screen tend to be slightly overestimated,

whereas y-coordinates at the bottom of the screen tend to be slightly underestimated (see also Figure 5b for a greater breakdown).

The average magnitude of the webcam error relative to the Eyelink was 0.10° (*SD*= 3.26°) for the x-dimension and 0.35° (*SD*= 2.32°) for the y-dimension (see Figure 4b). This again shows little systematic bias, but a somewhat large variance of the error distribution. Notably, the y-dimension was estimated with much less error and much less variance compared to Experiment 1. This suggests that the set-up in Experiment 2 (using a laptop and only 2 lines of text) led to more accurate tracking of y-positions.

2.2.2. Analysis of target word lexical frequency effect

Table 6 shows the means and standard deviations for the four fixation time measures on target words in Experiment 2, while Table 7 shows the corresponding model estimates.

Table 6

Mean first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on target words in Experiment 2 by eye-tracker and word frequency condition.

		FFD		SFD		GD		TVT	
Tracker	Frequency	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Webcam	low	236	119	239	120	249	125	293	173
Webcam	high	219	104	222	105	236	119	277	192
Eyelink	low	160	67	163	69	183	87	210	114
Eyelink	high	153	59	155	59	175	84	203	122

Note: *FFD* = first fixation duration; *SFD* = single fixation duration; *GD* = gaze duration; *TVT* = total viewing time; SD = standard deviation. All means and standard deviations are reported in milliseconds (ms).

Table 7

Coefficient estimates and 95% credible intervals (in square brackets) from Bayesian linear mixed models fitted to first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on target words in Experiment 2.

	FFD	SFD	GD	TVT
Intercept	**5.163**	**5.188**	**5.243**	**5.353**
	[5.105, 5.221]	**[5.118, 5.257]**	**[5.164, 5.321]**	**[5.254, 5.451]**
Frequency	**0.060**	**0.057**	**0.060**	**0.070**
	[0.021, 0.099]	**[0.015, 0.099]**	**[0.013, 0.108]**	**[0.014, 0.126]**
Tracker	**0.359**	**0.362**	**0.317**	**0.314**
	[0.298, 0.422]	**[0.298, 0.427]**	**[0.264, 0.371]**	**[0.250, 0.379]**
Frequency × Tracker	0.033	0.028	0.014	0.032
	[-0.018, 0.084]	[-0.024, 0.080]	[-0.039, 0.065]	[-0.028, 0.092]

Note: Effects for which the 95% CrI excludes 0 are highlighted in bold.

As in Experiment 1, we observed a frequency effect in the expected direction across both devices: low-frequency words elicited longer fixations than high-frequency words (FFD: $b = 0.06$, 95% CrI $[0.02,\ 0.10]$, $P(b > 0) = 1.00$, $ER_{b>0} = 550.72$, $BF_{10} = 1.51$; SFD: $b = 0.06$, 95% CrI $[0.01,\ 0.10]$, $P(b > 0) = 1.00$, $ER_{b>0} = 218.18$, $BF_{10} = 0.70$; GD: $b = 0.06$, 95% CrI $[0.01,\ 0.11]$, $P(b > 0) = 0.99$, $ER_{b>0} = 155.86$, $BF_{10} = 0.50$; TVT:

$b = 0.07$, 95% CrI $[0.01,\ 0.13]$, $P(b > 0) = 0.99$, $ER_{b>0} = 124.98$, $BF_{10} = 0.60$). Fixation times were again longer with the Webcam than with the Eyelink (FFD: $b = 0.36$, 95% CrI $[0.30,\ 0.42]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; SFD: $b = 0.36$, 95% CrI $[0.30,\ 0.43]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; GD: $b = 0.32$, 95% CrI $[0.26,\ 0.37]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; TVT: $b = 0.31$, 95% CrI $[0.25,\ 0.38]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$). Again, we did not observe a credible Frequency × Tracker interaction (FFD: $b = 0.03$, 95% CrI $[-0.02,\ 0.08]$, $P(b > 0) = 0.90$, $ER_{b>0} = 8.96$, $BF_{10} = 0.06$; SFD: $b = 0.03$, 95% CrI $[-0.02,\ 0.08]$, $P(b > 0) = 0.86$, $ER_{b>0} = 6.05$, $BF_{10} = 0.05$; GD: $b = 0.01$, 95% CrI $[-0.04,\ 0.07]$, $P(b > 0) = 0.69$, $ER_{b>0} = 2.27$, $BF_{10} = 0.03$; TVT: $b = 0.03$, 95% CrI $[-0.03,\ 0.09]$, $P(b > 0) = 0.86$, $ER_{b>0} = 6.20$, $BF_{10} = 0.05$).

Similar to Experiment 1, we summarized the low vs. high frequency contrast separately within each device using estimated marginal means. Interestingly, in Experiment 2, the Webcam data contained stronger evidence in favour of the frequency effect than the Eyelink data, with the 95% CI for the frequency effect excluding 0 for both devices in FFD (Webcam: *M* = 0.08 95% *CI* = [0.03, 0.13]; Eyelink: *M* = 0.04 95% *CI* = [0.00, 0.09]) and GD (Webcam: *M* = 0.07 95% *CI* = [0.01, 0.12]; Eyelink: *M* = 0.05 95% *CI* = [-0.00, 0.11]), although the exclusion was very narrow for the Eyelink data), but only for the Webcam data in SFD (Webcam: *M* = 0.07 95% *CI* = [0.02, 0.12]; Eyelink: *M* = 0.04 95% *CI* = [-0.00, 0.09]) and TVT (Webcam: *M* = 0.09 95% *CI* = [0.02, 0.15]; Eyelink: *M* = 0.05 95% *CI* = [-0.01, 0.12]). These somewhat unexpected results show that the Webcam data, although noisier, do not necessarily always underestimate the size of an effect. In any case, our results highlight that a large number of observations are needed to detect even strong effects such as

the word frequency effect reliably at a sampling rate of 30 Hz (as suggested by Angele et al., 2025).

2.2.3. Analysis of corpus lexical frequency

For Experiment 2, we also fit models predicting fixation times on all the words in every sentence (except for sentence-initial and sentence-final words), with Zipf frequency as a continuous predictor and tracker as a discrete factor with two levels, Eyelink and Webcam, as was done for the Experiment 1 data. Table 8 summarizes the Zipf frequency effect, the effect of the tracker used, and their interaction over all eligible words. We again observed a frequency effect, with higher frequency words being associated with shorter fixations (FFD: $b = -0.03$, 95% CrI $[-0.04,\ -0.02]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$; SFD: $b = -0.04$, 95% CrI $[-0.05,\ -0.02]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$; GD: $b = -0.06$, 95% CrI $[-0.08,\ -0.05]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$; TVT: $b = -0.09$, 95% CrI $[-0.10,\ -0.07]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$). Just as we had observed in all the previous analyses, the Webcam data resulted in longer detected fixations overall (FFD: $b = 0.44$, 95% CrI $[0.39,\ 0.50]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; SFD: $b = 0.44$, 95% CrI $[0.38,\ 0.49]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; GD: $b = 0.39$, 95% CrI $[0.35,\ 0.44]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$; TVT: $b = 0.36$, 95% CrI $[0.32,\ 0.41]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$). There was no credible evidence for an interaction between Zipf frequency and tracker in FFD ($b = 0.0096$, 95% CrI $[-0.0025,\ 0.02]$, $P(b > 0) = 0.94$, $ER_{b>0} = 15.21$, $BF_{10} = 0.02$), but there was a credible interaction in SFD ($b = 0.01$, 95% CrI $[0.0014,\ 0.03]$, $P(b > 0) = 0.99$, $ER_{b>0} = 66.51$,

$BF_{10} = 0.07$), GD ($b = 0.03$, 95% CrI $[0.02,\ 0.04]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} > 1000$), and TVT ($b = 0.03$, 95% CrI $[0.01,\ 0.04]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} = 11.10$). Importantly, the direction of this interaction was opposite to those observed in Experiment 1: Just as in the target word analysis, the frequency effect was stronger in the Webcam data compared to the Eyelink data. Finally, we again report `emtrends` estimates and 95% CIs for Eyelink and Webcam data separately. Just like Experiment 1, the frequency effect was credible for both the Eyelink and the Webcam data in all measures (FFD: Eyelink: *M* = -0.03 95% *CI* = [-0.05, -0.02]; Webcam: *M* = -0.02 95% *CI* = [-0.04, -0.01]; SFD: Eyelink: *M* = -0.04 95% *CI* = [-0.06, -0.03]; Webcam: *M* = -0.03 95% *CI* = [-0.05, -0.01]; GD: Eyelink: *M* = -0.08 95% *CI* = [-0.09, -0.06]; Webcam: *M* = -0.05 95% *CI* = [-0.06, -0.03]; TVT Eyelink: *M* = -0.10 95% *CI* = [-0.12, -0.08]; Webcam: *M* = -0.07 95% *CI* = [-0.09, -0.05]).

Table 8

Coefficient estimates and 95% credible intervals (in square brackets) from Bayesian linear mixed models fitted to first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on all words (except for the first and last in each sentence) in Experiment 2.

	FFD	SFD	GD	TVT
Intercept	**5.166**	**5.183**	**5.244**	**5.344**
	[**5.113, 5.219**]	[**5.124, 5.241**]	[**5.178, 5.311**]	[**5.269, 5.424**]
Zipf frequency	**-0.029**	**-0.037**	**-0.061**	**-0.085**
	[**-0.041, -0.018**]	[**-0.051, -0.023**]	[**-0.077, -0.046**]	[**-0.103, -0.067**]
Tracker	**0.444**	**0.436**	**0.395**	**0.364**
	[**0.390, 0.503**]	[**0.383, 0.489**]	[**0.348, 0.442**]	[**0.316, 0.411**]
Frequency × Tracker	0.010	**0.014**	**0.032**	**0.026**
	[-0.002, 0.022]	[**0.001, 0.026**]	[**0.021, 0.044**]	[**0.014, 0.037**]

Note: Effects for which the 95% CrI excludes 0 are highlighted in bold.

2.2.4. Summary and conclusions

To summarise, while Experiment 2 replicated many of the key findings from Experiment 1, the data also showed stronger effects in some cases. First, the webcam and Eyelink data exhibited similar, bit slightly higher, positive correlations. This again suggests that gaze

coordinates recorded by the webcam are broadly similar to those recorded by the Eyelink. Nevertheless, similar to Experiment 1, the webcam data still showed considerable variance of the error distributions, though there was little systematic bias and the y-dimension was estimated more accurately compared to Experiment 1. Therefore, these results suggest that the set-up employed in Experiment 2 (where the stimuli were presented on a laptop screen and were displayed over two lines of text) results in more accurate webcam tracking.

With regards to the reading measures, the data was again stronger, in the sense that the target word lexical frequency effect was now credibly detected also in the webcam data. Surprisingly, the webcam data showed numerically larger effects than the Eyelink data, though the means were also estimated with much more variance. In the corpus frequency analysis, both the webcam and Eyelink showed reliable frequency effects, replicating Experiment 1. Therefore, Experiment 2 showed similar, though stronger frequency effects in the webcam data.

Taken together, Experiments 1 and 2 show that frequency effects can *in principle* be detected with webcams under lab conditions. However, it is not clear if similar results can also be obtained in the "real world" where the experimenter has less control over the testing protocol and participants have webcams with a lower sampling resolution (typically 30 Hz vs. the 60Hz used in the lab studies). To help answer this question, we repeated Experiment 1 online, with participants completing the study at home[5].

[5] Experiment 3 still employed the same set-up of presenting the lexical frequency items over 3-4 lines of text. Because Experiments 2 and 3 were running largely in parallel, we could not use the previous findings to inform our design decisions. However, the advantage of this is that Experiment 3 provides a direct replication of Experiment 1 in the real world.

3. Experiment 3

3.1. Method

Sixty-six participants (32 female, 33 male, 1 preferred not to say) were recruited from Prolific.com for a payment of £10/hour. Participants reported having English as their native language, normal or corrected-to-normal vision (with contact lenses), and no prior history of reading disorders. Participants' average age was 32.04 years (*SD*= 7.86; range: 18-50). Participants' self-reported ethnicity was: White (65.15%), Black (21.21%), Asian (10.61%), and mixed (3.03%). Participants completed the same study as Experiment 1 at home. The first 18 participants were run as a pilot. They completed only the lexical frequency corpus task (the single-line corpus was used to pilot another task, which was not included in the final study). Therefore, the data for the single-sentence corpus was based only on 48 participants, while all 66 participants completed the lexical frequency corpus. The average effective sampling rate in the study was 25.66 Hz (SD= 6.70 Hz; range= 9.66- 35.66 Hz)[6]. The rest of the Method was identical to Experiment 1.

The data analysis was identical to the previous experiments. However, because participants completed the study at home, no Eyelink data is available. Therefore, the statistical models contained only Frequency as the predictor. The only difference to previous experiments is

[6] The webcam interface does not record the sampling rate of the webcam that participants had. However, based on the effective sampling rate (calculated on the raw data), it can be deduced that almost all participants had a 30 Hz web camera. The effective sampling rate is a bit lower than the expected rate of 30 Hz. While this is partly due to the software not achieving the full sampling rate in general (which was also evident in Experiments 1-2), the greater variance in effective sampling rate between participants may be due to differences in internet speed, which can affect the speed at which samples are sent to the server.

that the saccade detection threshold λ of Engbert and Kliegl's (2003) algorithm was reduced from 6 to 3 to deal with the lower sampling rate of the data (30 Hz). Using a threshold of λ=6 led to fewer and excessively long fixations; reducing it to λ=3 improved the sensitivity of the algorithm and led to fixation durations that were more similar to Experiments 1 and 2.

Samples that contained blinks (4.91%) or were otherwise outside the screen area (1.36%) were removed from analysis. Fixations shorter than 80 ms that were within 1 character distance of a temporally adjacent fixation were merged into that adjacent fixation (0.59%); any remaining fixations under 80 ms were discarded as outliers (6.94%). Additionally, fixations longer than 1000 ms were also discarded as outliers (5.78%).

Data analysis. The model specifications and data analysis were the same as in Experiment 1, except that Tracker was not included as a predictor since all data were collected remotely using the participants' webcam.

3.2. Results and Discussion

Participants' comprehension accuracy was 95.9% (*SD*= 19.6%; *range*: 85-100%), showing that they read for understanding. The average trial duration was 3864 ms (*SD*= 2801 ms).

3.2.1. Analysis of target word lexical frequency effect

Table 9 shows the mean and standard deviation of first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on target words in Experiment 3 by word frequency condition, while Table 10 shows the model estimates.

In Experiment 3, low-frequency target words elicited longer fixation times than high-frequency target words in FFD ($b = 0.07$, 95% CrI $[0.02,\ 0.13]$, $P(b > 0) = 0.99$, $ER_{b>0} = 180.82$, $BF_{10} = 0.70$), SFD ($b = 0.09$, 95% CrI $[0.03,\ 0.15]$, $P(b > 0) = 1.00$, $ER_{b>0} = 279.70$, $BF_{10} = 1.27$), GD ($b = 0.11$, 95% CrI $[0.05,\ 0.17]$, $P(b > 0) = 1.00$, $ER_{b>0} > 1000$, $BF_{10} = 5.34$), and TVT ($b = 0.09$, 95% CrI $[0.03,\ 0.16]$, $P(b > 0) = 1.00$, $ER_{b>0} = 252.97$, $BF_{10} = 1.22$). Overall, this demonstrates that, despite the much lower control we had over devices and experimental setup in the online study, reliable word frequency effects could still be found.

Table 9

Mean first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on target words in Experiment 3 by eye-tracker and word frequency condition.

	FFD		SFD		GD		TVT	
Frequency	Mean	SD	Mean	SD	Mean	SD	Mean	SD
low	381	226	385	227	436	279	521	372
high	363	226	359	226	402	272	493	386

Note: *FFD* = first fixation duration; *SFD* = single fixation duration; *GD* = gaze duration; *TVT* = total viewing time; SD = standard deviation. All means and standard deviations are reported in milliseconds (ms)

Table 10

Coefficient estimates and 95% credible intervals (in square brackets) from Bayesian linear mixed models fitted to first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on target words in Experiment 3.

	FFD	SFD	GD	TVT
Intercept	**5.760**	**5.752**	**5.851**	**5.987**
	[**5.681, 5.842**]	[**5.672, 5.835**]	[**5.773, 5.933**]	[**5.910, 6.064**]
Frequency	**0.074**	**0.090**	**0.110**	**0.095**
	[**0.018, 0.129**]	[**0.026, 0.154**]	[**0.047, 0.172**]	[**0.027, 0.161**]

Note: Effects for which the 95% CrI excludes 0 are highlighted in bold.

3.2.3. Analysis of corpus lexical frequency

As in Experiment 1, we also fitted Bayesian linear mixed models including all words in every sentence (both the multi-line sentences with the frequency manipulation and the single-line sentences) into the analysis. This increased the number of observations substantially compared to the target word analysis.

Table 11 shows the results of these “corpus-analysis style” Bayesian linear mixed models with Zipf frequency as a continuous predictor for the four fixation time measures on all words in the sentences (except for the first and last word in each sentence) in Experiment 3.

In Experiment 3, we observed a consistent effect of zipf frequency on all measures (FFD: $b = -0.04$, 95% CrI $[-0.05, -0.03]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$; SFD: $b = -0.04$, 95% CrI $[-0.05, -0.03]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$; GD: $b = -0.06$, 95% CrI $[-0.07, -0.05]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$; TVT: $b = -0.08$, 95% CrI $[-0.09, -0.06]$, $P(b < 0) = 1.00$, $ER_{b<0} > 1000$, $BF_{10} > 1000$). This again demonstrates that, when there is a sufficiently high number of observations, the frequency effect can be detected using online Webcam eye tracking, even though we did not have direct control over the experimental setup and the sampling rate of participants' cameras was generally lower.

Table 11

Coefficient estimates and 95% credible intervals (in square brackets) from Bayesian linear mixed models fitted to first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), and total viewing time (TVT) on all words (except for the first and last in each sentence) in Experiment 3.

	FFD	SFD	GD	TVT
Intercept	**5.656**	**5.656**	**5.726**	**5.820**
	[**5.588, 5.725**]	[**5.590, 5.723**]	[**5.660, 5.792**]	[**5.762, 5.878**]
Zipf frequency	**-0.036**	**-0.039**	**-0.059**	**-0.077**
	[**-0.047, -0.025**]	[**-0.051, -0.027**]	[**-0.071, -0.046**]	[**-0.090, -0.063**]

Note: Effects for which the 95% CrI excludes 0 are highlighted in bold.

3.2.3. Summary and conclusions

In summary, the eye-tracking data obtained online without any direct experimenter control of the testing environment were clearly noisier than the laboratory-based data. Despite this, we found reliable word frequency effects, both in terms of the categorical effect of the target word manipulation and in terms of the continuous effect of zipf frequency on all words. This shows that effects of cognitive processes on eye movements can be measured even in fully online eye-tracking studies if the sample size is large enough.

4. General Discussion

We performed three experiments to test whether it is feasible to study eye-movements during sentence reading using webcam-based eye-tracking as implemented in the Labvanced experimental platform. We were particularly interested in whether we could observe the word frequency effect, as this requires assigning fixations accurately to word-based regions of interest. In Experiments 1 and 2, we compared the performance of webcam eye-tracking as implemented in Labvanced with the SR Research EyeLink under controlled laboratory conditions. In Experiment 3, we recruited participants on an online platform to participate in an eye-tracking experiment using their own devices outside of the laboratory. In each of these experiments, we were able to find evidence for the word frequency effect in the webcam data, although the strength of the evidence varied by experimental setup and design. In the lab-based Experiments 1 and 2 that co-registered webcam and EyeLink data, we found that, as

expected, the webcam data were noisier than those recorded by the high-precision EyeLink eye-tracker. Despite this, there was a surprisingly high correlation between both data sources in terms of the estimated gaze position and, importantly, there was little systematic bias observed in the webcam data, especially in the x dimension, which is critically important for determining which word in a sentence a reader is fixating on. The relative absence of large directional biases means that we can reduce the amount of noise by averaging over multiple observations, compensating for the lower quality of the webcam data by increasing the sample size (similar to the findings by Angele et al., 2025, who found that the increased noise associated with tracking at lower sampling rates can be compensated by increasing the number of observations). Therefore, our results coincide with those of Kaduk et al. (2023), who also found high correlations between the gaze position estimates of the webcam and Eyelink in a free viewing task.

Of course, the focus of the current study was whether we can obtain useful information about fixation times on words in a reading task. The lower precision of the webcam data made it necessary to adapt the experimental design relative to how lab-based studies with high-precision eye-trackers are normally done. Instead of being able to present a long sentence using a medium to small font size in a single, vertically centered line, we used a much larger font size, either severely limiting the sentence length (such as in the single-line sentences presented in Experiments 1 and 3) or resulting in the sentences being split across multiple lines. We know that line breaks and the resulting return sweeps affect reading behavior (Parker et al., 2019), so this represents a limitation that cannot be easily addressed. On the other hand, a great deal of reading nowadays takes place on mobile devices with a limited

width and line length, so having shorter lines may not necessarily be unrepresentative of the reading process in general. Additionally, the assignment of fixations to lines in multi-line text is not trivial and can be the source of additional errors and/or require the application of correction algorithms (Carr et al., 2022; Mercier et al., 2024). In the present study, we did not apply any line correction algorithms and still obtained evidence of the frequency effect.

In Experiment 2, we attempted to strike a balance between sentence length, font size, and the ease of assigning fixations to lines by splitting each sentence into two lines and presenting them at the top and towards the bottom of the screen, respectively. This may be a satisfactory compromise when using webcam eye-tracking for medium-length sentences, although surely our design can be further refined in the future.

A strong correlation between webcam and EyeLink positional data is necessary, but not sufficient for generating similar fixation time measures, as these depend on the correct assignment of gaze position to word regions. A relatively large error that results in the fixation being assigned to the correct word has no impact on the results, while even a small error that moves the fixation to the wrong region can introduce great amounts of noise to the fixation time measurements. We found that, despite this concern, the EyeLink and the webcam data showed frequency effects in the same direction, even though these were not always statistically reliable in the noisier webcam data (e.g., Experiment 1). As described above, collecting more data can help with this issue: in the corpus-style analyses which included observations from most of the words in each sentence, the evidence for the frequency effect was much clearer in the webcam data than it was in the analyses of the experimental target word frequency manipulation. The latter involved a stronger

experimental control of the sentence context and target word properties but resulted in fewer observations as only one word per sentence could be analysed. This is not to say that experimental designs cannot be done using online eye-tracking, but researchers should plan to collect data from many participants and, ideally, use many items to reach a high number of observations per condition.

Overall, we observed clear evidence of the word frequency effect in all experiments. The target word frequency effect was numerically similar in all experiments in the webcam data (see Figure 7). While the target word frequency effect was not credible in Experiment 1, the corpus-level analysis nevertheless showed evidence for a corpus frequency effect. We therefore generally replicated the findings by Angele et al. (2025), who, based on downsampling high-precision EyeLink data, predicted that it should be possible to detect the word frequency effect even at sampling rates as low as 60 or 30 Hz. By finding the frequency effect in the webcam data, we confirmed their prediction that this would also be the case with non-downsampled data that were recorded at lower sampling rates.

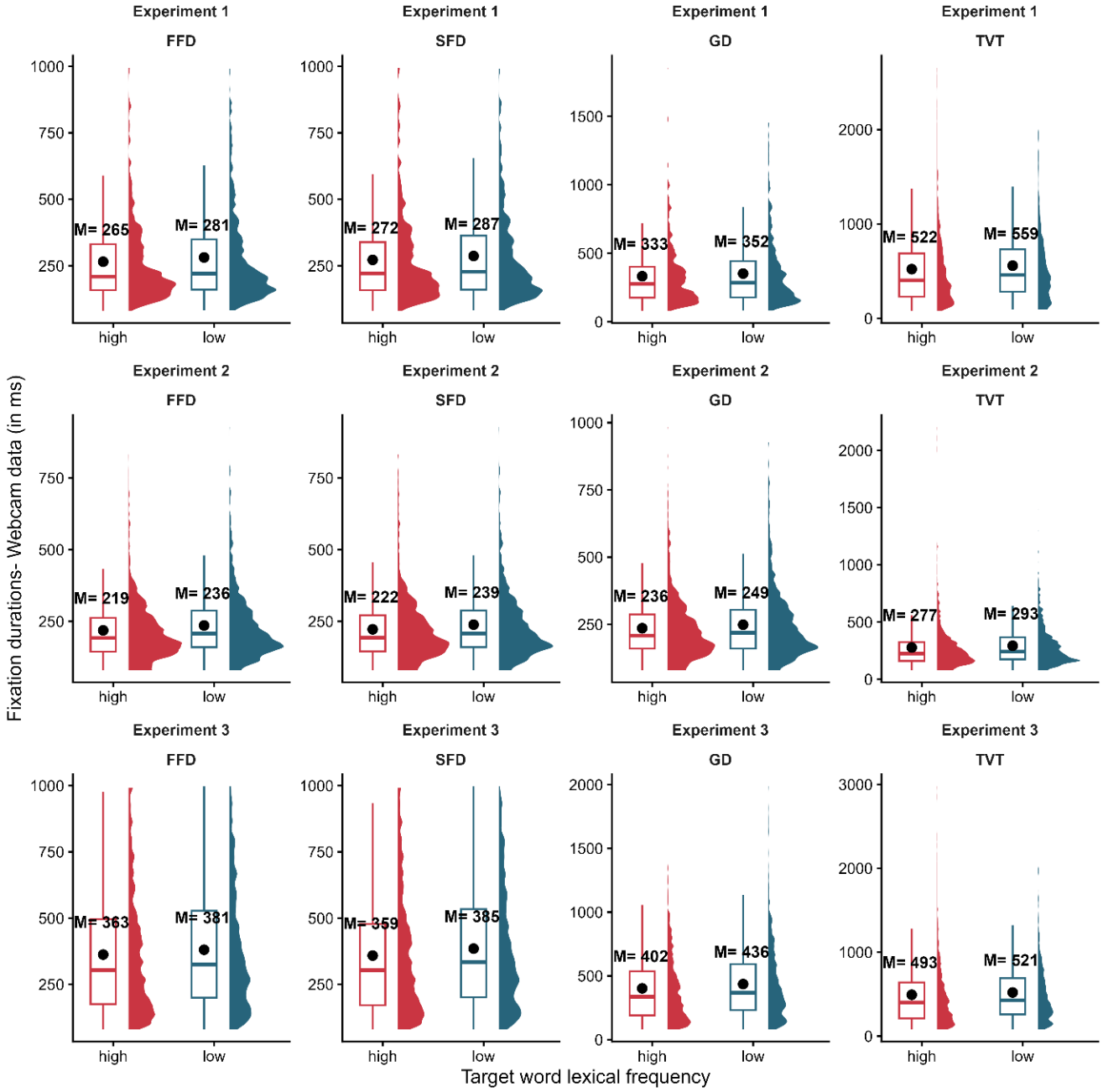

Figure 7. Fixation durations in the webcam data, showing the target word lexical frequency effect in Experiments 1-3. The graph shows the boxplots on the left and the density distributions on the right. Black dots show the mean, whereas the boxplots show the median.

An important finding from Experiment 2 is that the noisier webcam data do not always necessarily underestimate the effect size: while noise is likely to result in weaker evidence for an effect, sometimes the noise in the data can result in a larger effect size estimate. A larger sample size should also help with obtaining a more accurate effect size estimate by reducing variance around the means.

Experiment 3 demonstrated that running an eye-tracking experiment completely online, with all the variability this implies in terms of environment, device, participant compliance with experimenter instructions etc., does result in noisier data, but this did not prevent us from finding evidence for the word frequency effect. Researchers should plan to collect large samples of participants for online experiments, which of course is strongly facilitated given that an online study can test multiple participants at the same time (subject to platform limitations) and do not require the presence or attention of an experimenter to run.

Our findings demonstrate that webcam-based eye tracking, as implemented in the Labvanced platform, can successfully capture both spatial gaze patterns and word-level cognitive effects during sentence reading, including the classical word frequency effect. These results are consistent with, and extend, previous work showing that webcam eye tracking can detect robust cognitive effects when regions of interest are sufficiently large or well separated (Slim & Hartsuiker, 2023; Prystauka et al., 2023; Slim et al., 2024; Kandel & Snedeker, 2025). Similarly, Kaduk et al. (2023) reported high positional correlations between Labvanced and Eyelink data in attention tasks, but their study did not address reading; our findings extend this work by demonstrating that Labvanced can also detect word-level reading effects. Furthermore, our results provide empirical support for Angele et al.'s (2025) prediction that

lexical frequency effects should remain detectable even at sampling rates as low as 30–60 Hz. However, webcam eye-tracking is limited in its ability to provide the high spatial precision required for certain cognitive tasks involving fine-grained gaze discrimination (James et al., 2025), but our findings indicate that, when noise is mitigated through increased observations and appropriate design adaptations, webcam-based tracking can nevertheless be adequate for detecting word-level effects during sentence reading.

In the future, more experimental platforms and experimental designs should be evaluated. For example, would it be possible to record participants reading very short paragraphs with multiple sentences in an online experiment and still obtain useful data? Given its lower precision, we are unsure if our experimental approach would work with Webgazer (Papoutsaki et al., 2016), but it is possible that useful information could be obtained with very large text and a very large sample of participants. We recommend that researchers who wish to use a different platform for their studies do a pilot study first, demonstrating the feasibility of detecting a well-established effect such as the word frequency effect. If the size of the effect of interest is smaller than the size of the word frequency effect observed in the present study, online eye-tracking may not be an adequate technique to use. Nevertheless, we encourage researchers to investigate statistical power in the context of online eye-tracking during reading and to attempt to provide an estimate of the smallest effect sizes that can be realistically detected using this technique.

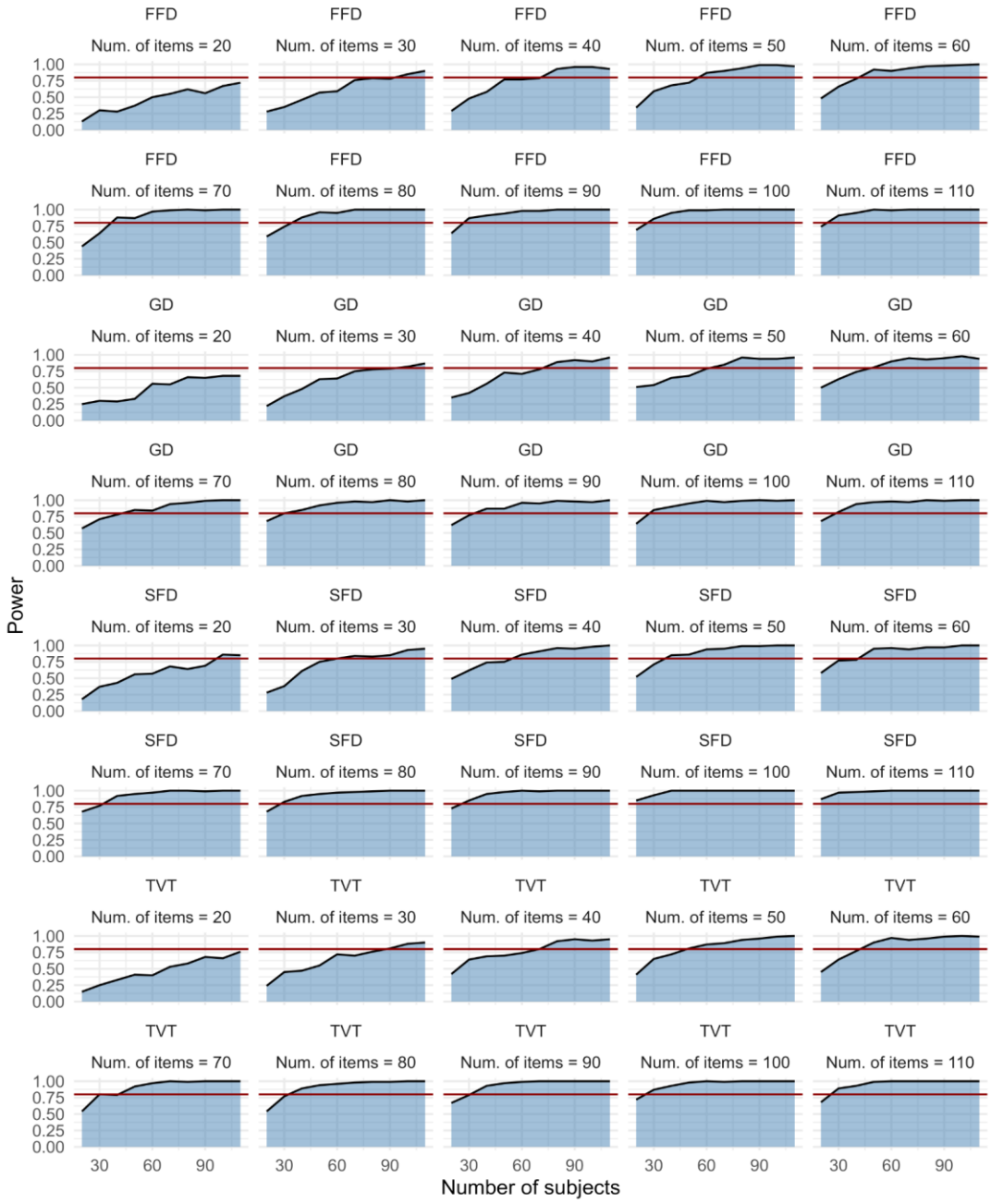

Figure 8. Statistical simulations showing the empirical power to detect the target word lexical frequency effect in the webcam data from Experiment 3. The results are broken down by different subject and item numbers, with the 80% power level indicated by a horizontal red line. Simulations were conducted with the *simr* package (Green & MacLeod, 2016). Power

was calculated with 100 simulations per cell and included 10% random data loss. The models included random intercepts for subjects and items, and random slopes for target word frequency.

To estimate power for the lexical frequency effect, we conducted statistical power simulations with the *simr* package (Green & MacLeod, 2016). Figure 8 shows the statistical power needed to detect the target word lexical frequency effect based on the webcam results from Experiment 3. The results showed that 60 participants and 60 items are enough to achieve 80% statistical power of detecting the lexical frequency effect in all measures. The number of participants can be offset by a larger number of items, and vice versa. However, researchers should do their own calculations for other effects.

We hope that our results will encourage other researchers to attempt eye-tracking experiments using webcams, while keeping in mind the sample size requirements. Importantly, webcam research does not necessarily have to be performed completely online: a researcher wishing to study eye-tracking and requiring a slightly higher precision than that available online might run a webcam experiment on a laptop or a computer with an external webcam, as we did in Experiments 1 and 2. By using an external webcam, the quality of the data may be further enhanced. Both online and in the laboratory, our results make us optimistic about the potential of webcam eye tracking being employed to help the study of reading in places and with populations where high-precision eye-tracker access is difficult.

Conclusion

In summary, our results clearly demonstrate that it is possible to obtain useful eye-tracking data about the processing of individual words during sentence reading in an online experiment, at least using the experimental design and the online platform (Labvanced) that we employed. We hope that webcam eye-tracking will complement lab-based eye-trackers as an additional method to obtain useful data about language processing during reading, particularly in difficult-to-reach populations.

Acknowledgements: Scicovery GmbH provided us with access to Labvanced.com free of charge for the purpose of collecting this data. Apart from granting us access to Labvanced.com, neither Scicovery GmbH nor any of its employees were involved in the conceptualisation, design, data acquisition, data analysis, or the writing of this manuscript. This paper should not be taken as an endorsement of any products or services. We thank the students in module PSYC0012 at UCL (2024/2025), who did a mini project on the topic and assisted with part of the data collection on Prolific.com for Experiment 3. We also thank Lorena Chicharro for her help with data collection for Experiment 2.

Data availability: The data, code, and materials from this study are available at https://osf.io/axq7u/

Conflict of interest statement: None of the authors are affiliated with or have any financial interest in Scicovery GmbH or Labvanced.com. The authors received access to Labvanced.com free of charge for the purpose of collecting the data, but there was no other involvement of Scicovery GmbH or any of its employees in any aspect of design, data collection, analysis, or writing of this paper.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Andersson, R., Nyström, M., & Holmqvist, K. (2010). Sampling frequency and eye-tracking measures: How speed affects durations, latencies, and more. *Journal of Eye Movement Research*, *3*(3), Article 3. https://doi.org/10.16910/jemr.3.3.6

Angele, B., Baciero, A., Gómez, P., & Perea, M. (2022). Does online masked priming pass the test? The effects of prime exposure duration on masked identity priming. *Behavior Research Methods*, *55*(1), 151–167. https://doi.org/10.3758/s13428-021-01742-y

Angele, B., & Duñabeitia, J. A. (2024). Closing the eye-tracking gap in reading research. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1425219

Angele, B., Gunes Ozkan, Z., Serrano-Carot, M., & Duñabeitia, J. A. (2025). How low can you go? Tracking eye movements during reading at different sampling rates. *Behavior Research Methods*, *57*(7), 195. https://doi.org/10.3758/s13428-025-02713-3

Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01501-5

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Arel-Bundock, V. (2022). modelsummary: Data and Model Summaries in *R*. *Journal of*

Statistical Software, *103*(1). https://doi.org/10.18637/jss.v103.i01

Aust, F., & Barth, M. (2022). *papaja: Prepare American Psychological Association Journal Articles with R Markdown* (p. 0.1.4) [Data set]. https://doi.org/10.32614/CRAN.package.papaja

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barth, M. (2025). *tinylabels: Lightweight variable labels* [Computer software]. https://doi.org/10.32614/CRAN.package.tinylabels

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Bramlett, A. A., & Wiener, S. (2025). The art of wrangling: Working with web-based visual world paradigm eye-tracking data in language research. *Linguistic Approaches to Bilingualism*, *15*(4), 538–570. https://doi.org/10.1075/lab.23071.bra

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Bürkner, P.-C. (2017). brms: An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian Item Response Modeling in *R* with brms and *Stan*. *Journal of Statistical Software*, *100*(5). https://doi.org/10.18637/jss.v100.i05

Carr, J. W., Pescuma, V. N., Furlan, M., Ktori, M., & Crepaldi, D. (2022). Algorithms for the automated correction of vertical drift in eye-tracking data. *Behavior Research Methods*, *54*(1), 287–310. https://doi.org/10.3758/s13428-021-01554-0

Ching, T. (2025). *qs2: Efficient serialization of r objects* [Computer software]. https://github.com/qsbase/qs2

Choi, W., Desai, R. H., & Henderson, J. M. (2014). The neural substrates of natural reading: A comparison of normal and nonword text using eyetracking and fMRI. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.01024

Clifton, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40-year legacy. *Journal of Memory and Language*, *86*, 1–19. https://doi.org/10.1016/j.jml.2015.07.004

Cornsweet, T. N., & Crane, H. D. (1973). Accurate two-dimensional eye tracker using first and fourth Purkinje images. *JOSA*, *63*(8), 921–928. https://doi.org/10.1364/JOSA.63.000921

Degno, F., Loberg, O., & Liversedge, S. P. (2021). Co-Registration of Eye Movements and Fixation—Related Potentials in Natural Reading: Practical Issues of Experimental Design and Data Analysis. *Collabra: Psychology*, *7*(1). https://doi.org/10.1525/collabra.18032

Dickey, J. M., & Lientz, B. P. (1970). The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226. https://doi.org/10.1214/aoms/1177697203

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, *20*(5), 917–930. https://doi.org/10.1017/S1366728916000547

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General*, *140*(4), 552–572. https://doi.org/10.1037/a0023885

Dowle, M., & Srinivasan, A. (2021). *data.table: Extension of `data.frame`*. https://CRAN.R-project.org/package=data.table

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, *45*(4), 1246-1258.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. https://doi.org/10.1016/S0042-6989(03)00084-1

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies. *International Conference on*

Computational Social Science (Cologne), 1–3. https://www.researchgate.net/profile/Caspar-Goeke/publication/322273524_LabVanced_A_Unified_JavaScript_Framework_for_Online_Studies/links/5a4f7ac64585151ee284d8c2/LabVanced-A-Unified-JavaScript-Framework-for-Online-Studies.pdf

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, *41*(1), 163–171. https://doi.org/10.3758/BRM.41.1.163

Godwin, H. J., Lee, C. E., & Drieghe, D. (2025). A multiverse analysis of cleaning and analyzing procedures of eye movement data during reading. *Behavior Research Methods*, *57*(6), 164. https://doi.org/10.3758/s13428-025-02689-0

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, *40*(3). https://doi.org/10.18637/jss.v040.i03

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). lab.js: A free, open, online study builder. *Behavior Research Methods*, *54*(2), 556–573. https://doi.org/10.3758/s13428-019-01283-5

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study

language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

Huey, E. B. (with University of California Libraries). (1908). *The psychology and pedagogy of reading, with a review of the history of reading and writing and of methods, texts, and hygiene in reading*. New York : Macmillan. http://archive.org/details/psychologypedago00hueyiala

Hutt, S., Wong, A., Papoutsaki, A., Baker, R. S., Gold, J. I., & Mills, C. (2023). Webcam-based eye tracking to detect mind wandering and comprehension errors. *Behavior Research Methods*, *56*(1), 1–17. https://doi.org/10.3758/s13428-022-02040-x

Iannone, R., Cheng, J., Schloerke, B., Hughes, E., Lauer, A., Seo, J., Brevoort, K., & Roy, O. (2025). Gt: Easily create presentation-ready display tables. *(No Title)*. https://gt.rstudio.com

James, A. N., Ryskin, R., Hartshorne, J. K., Backs, H., Bala, N., Barcenas-Meade, L., Bhattarai, S., Charles, T., Copoulos, G., Coss, C., Eisert, A., Furuhashi, E., Ginell, K., Guttman-McCabe, A., Harrison, E. (Chaz), Hoban, L., Hwang, W. A., Iannetta, C., Koenig, K. M., … De Leeuw, J. R. (2025). What Paradigms Can Webcam Eye-Tracking Be Used For? Attempted Replications of Five Cognitive Science Experiments. *Collabra: Psychology*, *11*(1), 140755. https://doi.org/10.1525/collabra.140755

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*, *56*(5), 5002–5022. https://doi.org/10.3758/s13428-023-02237-8

Kandel, M., & Snedeker, J. (2025). Assessing two methods of webcam-based eye-tracking for child language research. *Journal of Child Language*, *52*(3), 675–708. https://doi.org/10.1017/S0305000924000175

Kay, M. (2024). *tidybayes: Tidy Data and Geoms for Bayesian Models* (Version v3.0.7) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.1308151

Kukona, A., Cho, P. W., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(2), 326–347. https://doi.org/10.1037/a0034903

Kuperman, V., Schroeder, S., Acartürk, C., Agrawal, N., Alexandre, D. M., Bolliger, L. S., Brasser, J., Campos-Rojas, C., Drieghe, D., Đurđević, D. F., Freitas, L. V. G. de, Goldina, S., Orellana, R. I., Jäger, L. A., Jóhannesson, Ó. I., Khare, A., Kharlamov, N., Knudsen, H. B. S., Kristjánsson, Á., … Siegelman, N. (2025). New data on text reading in English as a second language: The Wave 2 expansion of the Multilingual Eye-Movement Corpus (MECO). *Studies in Second Language Acquisition*, *47*(2), 677–695. https://doi.org/10.1017/S0272263125000105

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lenth, R. V., & Piaskowski, J. (2025). *Emmeans: Estimated marginal means, aka least-squares means* [Computer software]. https://rvlenth.github.io/emmeans/

Ludbrook, J. (2010). Confidence in Altman–Bland plots: A critical review of the method of differences. *Clinical and Experimental Pharmacology and Physiology*, *37*(2), 143–

149. https://doi.org/10.1111/j.1440-1681.2009.05288.x

Lüdecke, D. (2025). *sjPlot: Data visualization for statistics in social science* [Computer software]. https://CRAN.R-project.org/package=sjPlot

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, *81*(12), 899–917.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, *17*, 578–587.

Mercier, T. M., Budka, M., Vasilev, M. R., Kirkby, J. A., Angele, B., & Slattery, T. J. (2024). Dual input stream transformer for vertical drift correction in eye-tracking reading data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–12. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2024.3411938

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*(5), 368–378. https://doi.org/10.1016/j.jmp.2011.06.004

Müller, K. (2025). *Here: A simpler way to find your files* [Computer software]. https://here.r-lib.org/

Müller, K., & Wickham, H. (2025). *Tibble: Simple data frames* [Computer software]. https://tibble.tidyverse.org/

Pan, Y., Frisson, S., & Jensen, O. (2021). Neural evidence for lexical parafoveal processing. *Nature Communications*, *12*(1), 5234. https://doi.org/10.1038/s41467-021-25571-x

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable Webcam Eye Tracking Using User Interactions. *IJCAI'16:*

Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 3839–3845. https://dl.acm.org/doi/10.5555/3061053.3061156

Parker, A. J., Slattery, T. J., & Kirkby, J. A. (2019). Return-sweep saccades during reading in adults and children. *Vision Research*, *155*, 35–43. https://doi.org/10.1016/j.visres.2018.12.007

Patterson, A. S., Nicklin, C., & Vitta, J. P. (2025). Methodological recommendations for webcam-based eye tracking: A scoping review. *Research Methods in Applied Linguistics*, *4*(3), 100244. https://doi.org/10.1016/j.rmal.2025.100244

Peirce, J., Hirst, R., & MacAskill, M. (2022). *Building Experiments in PsychoPy*. SAGE.

Prystauka, Y., Altmann, G. T. M., & Rothman, J. (2023). Online eye tracking and real-time sentence processing: On opportunities and efficacy for capturing psycholinguistic effects of different magnitudes and diversity. *Behavior Research Methods*, *56*(4), 3504–3522. https://doi.org/10.3758/s13428-023-02176-4

R Core Team. (2025). *R: A language and environment for statistical computing. R Foundation for Statistical Computing* [Computer software]. https://www.R-project.org/

Rayner, K. (1975). Parafoveal identification during a fixation in reading. *Acta Psychologica*, *39*(4), 271–281. https://doi.org/10.1016/0001-6918(75)90011-6

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). The Thirty-Fifth Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of*

Experimental Psychology, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201. https://doi.org/10.3758/BF03197692

Ribeiro, T., Brandl, S., Søgaard, A., & Hollenstein, N. (2024). *WebQAmGaze: A Multilingual Webcam Eye-Tracking-While-Reading Dataset* (No. arXiv:2303.17876). arXiv. https://doi.org/10.48550/arXiv.2303.17876

Richardson, D. C., & Spivey, M. J. (2008). Eye Tracking: Characteristics and Methods. In *Encyclopedia of Biomaterials and Biomedical Engineering* (2nd edn). CRC Press.

Ryskin, R. A., Qi, Z., Duff, M. C., & Brown-Schmidt, S. (2017). Verb biases are shaped through lifelong learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 781–794. https://doi.org/10.1037/xlm0000341

Saxena, S., Fink, L. K., & Lange, E. B. (2023). Deep learning models for webcam eye tracking in online experiments. *Behavior Research Methods*, *56*(4), 3487–3503. https://doi.org/10.3758/s13428-023-02190-6

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, *74*(1), 5–35. https://doi.org/10.3758/s13414-011-0219-2

Shannon, C. E. (1949). Communication in the Presence of Noise. *Proceedings of the IRE*, *37*(1), 10–21. Proceedings of the IRE. https://doi.org/10.1109/JRPROC.1949.232969

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Da Fonseca, S. M., Dirix, N., Duyck,

W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Kwon, N., Lõo, K., … Kuperman, V. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, *54*(6), 2843–2863. https://doi.org/10.3758/s13428-021-01772-6

Slim, M. S., & Hartsuiker, R. J. (2023). Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js. *Behavior Research Methods*, *55*(7), 3786–3804. https://doi.org/10.3758/s13428-022-01989-z

Slim, M. S., Kandel, M., Yacovone, A., & Snedeker, J. (2024). Webcams as Windows to the Mind? A Direct Comparison Between In-Lab and Web-Based Eye-Tracking Methods. *Open Mind*, *8*, 1369–1424. https://doi.org/10.1162/opmi_a_00171

Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A., Franchin, L., Garbisch, I., Grosse Wiesmann, C., Hamlin, J. K., Havron, N., Hay, J. F., Hermansen, T. K., Jakobsen, K. V., Kalinke, S., Ko, E.-S., Kulke, L., Mayor, J., Meristo, M., … Schuwerk, T. (2024). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *Infancy*, *29*(1), 31–55. https://doi.org/10.1111/infa.12564

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus, Eye Movements, and Vision. *I-Perception*, *1*(1), 7–27. https://doi.org/10.1068/i0382

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. https://doi.org/10.1016/j.visres.2012.10.012

Thilderkvist, E., & Dobslaw, F. (2024). On current limitations of online eye-tracking to study

the visual processing of source code. *Information and Software Technology*, *174*, 107502. https://doi.org/10.1016/j.infsof.2024.107502

Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., & Navalpakkam, V. (2020). Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications*, *11*(1), 4553. https://doi.org/10.1038/s41467-020-18360-5

Van Boxtel, W. S., Linge, M., Manning, R., Haven, L. N., & Lee, J. (2024). Online Eye Tracking for Aphasia: A Feasibility Study Comparing Web and Lab Tracking and Implications for Clinical Use. *Brain and Behavior*, *14*(11), e70112. https://doi.org/10.1002/brb3.70112

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Vasilev, M. R., Liversedge, S. P., Rowan, D., Kirkby, J. A., & Angele, B. (2019). Reading is disrupted by intelligible background speech: Evidence from eye-tracking. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(11), 1484–1512. https://doi.org/10.1037/xhp0000680

von der Malsburg, T. (2019). *saccades: Detection of Fixations in Eye-Tracking Data* (Version 0.2-1) [Computer software]. https://github.com/tmalsburg/saccades

Wade, N. J. (2010). Pioneers of eye movement research. *I-Perception*, *1*(2), 33–68. https://doi.org/10.1068/i0389

Wade, N. J., & Tatler, B. W. (2008). Did Javal measure eye movements during reading?

Journal of Eye Movement Research, *2*(5), Article 5. https://doi.org/10.16910/jemr.2.5.5

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H. (2023). *Conflicted: An alternative conflict resolution strategy* [Computer software]. https://conflicted.r-lib.org/

Wickham, H. (2025a). *Forcats: Tools for working with categorical variables (factors)* [Computer software]. https://forcats.tidyverse.org/

Wickham, H. (2025b). *Stringr: Simple, consistent wrappers for common string operations* [Computer software]. https://stringr.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation* [Computer software]. https://dplyr.tidyverse.org

Wickham, H., & Henry, L. (2025). *Purrr: Functional programming tools* [Computer software]. https://purrr.tidyverse.org/

Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data* [Computer software]. https://readr.tidyverse.org

Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data* [Computer software]. https://tidyr.tidyverse.org

Xie, Y. (2015). *Dynamic Documents with R and knitr, Second Edition*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315382487

Yarbus, A. L. (1967). Eye Movements During Perception of Complex Objects. In A. L. Yarbus (Ed.), *Eye Movements and Vision* (pp. 171–211). Springer US. https://doi.org/10.1007/978-1-4899-5379-7_8