GLOSSA psycholinguistics

Vos, M., Minor, S., & Ramchand, G. (2022). Comparing infrared and webcam eye tracking in the Visual World Paradigm. *Glossa Psycholinguistics*, 1(1): 9, pp. 1–37. DOI: https://doi.org/10.5070/G6011131

eScholarship University of California

Comparing infrared and webcam eye tracking in the Visual World Paradigm

Myrte Vos, UiT the Arctic University of Norway, NO, myrte.vos@uit.no
Serge Minor, UiT the Arctic University of Norway, NO, sergey.minor@uit.no
Gillian Ramchand, UiT the Arctic University of Norway, NO, gillian.ramchand@uit.no

Visual World eye tracking is a temporally fine-grained method of monitoring attention, making it a popular tool in the study of online sentence processing. Recently, while infrared eye tracking was mostly unavailable, various web-based experiment platforms have rapidly developed webcam eye tracking functionalities, which are now in urgent need of testing and evaluation. We replicated a recent Visual World study on the incremental processing of verb aspect in English using 'out of the box' webcam eye tracking software (jsPsych; de Leeuw, 2015) and crowdsourced participants, and fully replicated both the offline and online results of the original study. We furthermore discuss factors influencing the quality and interpretability of webcam eye tracking data, particularly with regards to temporal and spatial resolution; and conclude that remote webcam eye tracking can serve as an affordable and accessible alternative to lab-based infrared eye tracking, even for questions probing the time-course of language processing.

1 Introduction

This paper presents a method study of webcam eye tracking as an alternative mode of collecting data in the Visual World Paradigm, a popular experimental paradigm in psycholinguistics. While webcam eye tracking technology has been under development and in use in some form or other for nearly a decade, two confluent factors have prompted several web-based experiment platforms creating eye tracking tools for behavioral research. One, major improvements to its accuracy and accessibility through the open-source browser tool WebGazer; and the other, the urgent need for a remote alternative to lab-based infrared eye tracking following the outbreak of COVID-19. We evaluate one of these tools, developed within the jsPsych library (version 6.3 De Leeuw, 2015), by replicating one of our own recent Visual World studies with a fully web-based experiment.

1.1 Webcam eye tracking

Most eye tracking systems used in behavioural research laboratories are infrared eye trackers: in brief, they project near-infrared light onto the pupils, which creates a corneal reflection (also known as a Purkinje image) that can be used to triangulate the visual angle of gaze. Eye trackers marketed towards scientists are often bundled together with proprietary stimuli presentation and data pre-processing software: though convenient and more user-friendly on the one hand, this can hinder the researcher in customizing the technology and experiment design beyond the options provided by the vendor, or accessing the raw data. Infrared eye trackers are also expensive, costing several thousand dollars at minimum; indeed, before the COVID-19 pandemic, innovation in eye tracking technology was mostly driven by a need to make it cheaper and more portable – not only for researchers, but for consumer-grade eye tracking applications and devices. As internet speed increased and crowdsourced workers became easily accessible through companies like Qualtrics and Amazon Mechanical Turk, the past decade saw the debut of many different ‘neuromarketing’ applications: e.g. Turkergaze (Xu et al., 2015), GazeParser (Sogo, 2013), WebGazer (Papoutsaki et al., 2016), GazeHawk, GazeRecorder, EyesDecide, RealEye, EyeSee, etc. Of these, WebGazer has emerged as the clear favourite for use in browser-based research in cognitive science.

Unlike most other eye tracking tools, WebGazer[1] maps eye features onto positions on the screen using dynamic, mouseclick-based calibration, taking advantage of the rule of thumb that users navigating a web page will look directly at where they click (Chen et al., 2001; Hauger et al., 2011; Huang et al., 2012). This reliance on natural browsing behaviour makes it better suited to User Interaction research than more ‘traditional’ behavioural research paradigms (Papoutsaki et al., 2017). However, WebGazer has several advantages that make it an attractive tool for

[1] The full name is WebGazer.js, also sometimes written as webgazer.js or WebGazer.js; in this paper we will refer to it simply as WebGazer.

behavioural scientists: it is fully integrated in the browser, without requiring users to download software; it computes and outputs gaze data in the form of [*x*,*y*,*t*] coordinates on the client browser, without transmitting video data to the experiment server; its design is modular, making it easy to substitute alternatives for the default facial recognition algorithm and ridge regression model; and the fact that it is free, open source, and actively maintained.

Semmelmann & Weigelt (2018) were the first to report a method study evaluating the usefulness of WebGazer in cognitive research. They conducted three common eye tracking tasks (fixation, pursuit, and free viewing) with custom-written experiment software integrating WebGazer, testing both in-lab and remote participants. They found an average spatial offset of, respectively, 15% (approx. 4° visual angle) and 18% of screen size, and an average saccade duration of 450 ms and 750 ms,[2] with significantly more variance in the remote sample. For comparison, a commercial infrared eye tracker sampling at >120 Hz can be expected to record saccade durations of 200 ms or less, with spatial offsets between 0.1° and 0.5° (Ehinger et al., 2019; Ooms et al., 2015).

However, despite the noisier, lower-resolution data, they were able to replicate a well-known eye tracking result: namely, that Western participants learning and categorizing human faces pay particular attention to the eye region (in contrast to participants from other cultural backgrounds) (Blais et al., 2008; and others). In the wake of their cautiously optimistic assessment, a handful of WebGazer-based experiments followed: Federico & Brandimonte (2019) used WebGazer in a lab setting through a commercial platform and a consumer-grade webcam; e.g. Yang & Krajbich (2021), Degen et al. (2021), and Madsen et al. (2021) integrated WebGazer into their own experiment code to run remote eye tracking experiments in the browser. Though their data were encouraging, generally replicating effects found with infrared eye tracking, they were also much noisier due to the differences in computer hardware, operating system, processing capacity, and lighting quality between participants. In addition, programming and hosting these experiments required considerable time, effort, and specialized skills. The incentive for cognitive scientists to invest in webcam eye tracking therefore remained low.

Since the outbreak of the COVID-19 pandemic, which largely precluded in-lab research and infrared eye tracking, several popular behavioural experiment software programs and libraries (at last count: PCIbex, Gorilla, jsPsych, and PsychoPy) have developed webcam eye tracking functionalities, most[3] of which rely on WebGazer. This, in tandem with a recent proliferation of researcher-friendly web hosting solutions (e.g. JATOS, Pushkin) and companies that

[2] This was not (as an anonymous reviewer supposed) a typo: Semmelmann & Weigelt (2018) use the word saccade to refer to the window of time during which their participants switch fixation targets, and measure its duration from the onset of the new fixation cross, to the moment the gaze "fully reached" the target.

[3] Not all; e.g. Labvanced (Finger et al., 2017) has its own proprietary eye tracking software; see e.g. Bánki et al. (2022) and Chouinard et al. (2019) for studies using LabVanced for infant eye tracking research.

combine experiment building graphical user interfaces and web hosting (e.g. Gorilla, Pavlovia, FindingFive), has made conducting webcam eye tracking experiments much more accessible. With the new wealth of possibilities comes the need to map out its caveats and limitations: it is already evident that dependent measures requiring very fine-grained temporal and spatial resolution, such as eye movements during reading, cannot be usefully investigated using webcam eye tracking. But the most fine-grained resolution at which it *can* be useful has not yet been pinned down, especially as the technology improves; and one experimental method where it almost certainly can at least supplement infrared eye tracking, is the Visual World Paradigm (Degen et al., 2021; Slim & Hartsuiker, 2021b).

1.2 The Visual World Paradigm

The Visual World Paradigm is one of the most productive methods in online language processing research, owing to the fact that human visual attention is tightly coupled with linguistic processing (Cooper, 1974). Given a 'visual world', i.e. a display of scenes or objects, and an auditory linguistic stimulus, participants' eye movements will gravitate towards those parts of the display that are associated in some way with what they hear (Tanenhaus et al., 1995; Allopenna et al., 1998; see Falk Huettig & Meyer, 2011 for review). These fixations are closely time-locked to the linguistic stimulus, often occurring before or within 200 ms of the target word's offset;[4] they have also been found to reflect predictive processing, in cases where the selectional restrictions of an earlier word constrain the possible targets in the visual display. In Altmann & Kamide (1999)'s eminent example, "The boy will eat the cake" triggered looks towards a cake (the only edible object in the display) *before* onset of the noun. There are several possible linking hypotheses for the relationship between eye movements and linguistic processing (see e.g. Falk Huettig & Meyer, 2011 and Magnuson, 2019 for discussion), and the formulation of a model integrating visual processing, linguistic processing, eye movement mechanics and high-level discourse and nonlinguistic cognitive factors is a priority for this paradigm (see e.g. Huettig et al., 2020; Chabal et al., 2022; Degen et al., 2021). For our purposes, it will suffice to say that Visual World Paradigm studies have shown, to quote Magnuson (2019), that "listeners are sensitive to every potentially useful (i.e., predictive) constraint that has been tested as early as we can measure." (p.134) The constraint that we investigated in the current studies is grammatical aspect: we give a theoretical motivation for this work in the section below, but readers who are interested primarily in the methodological results may take our word for it and proceed to section 2.

[4] 200 ms being the average latency of a saccade from one visual target to another (Saslow, 1967); though see e.g. Magnuson et al. (2008) and Huettig & Altmann (2011) for examples of how fixations can be delayed or suppressed depending on task conditions.

1.3 The original study

The experiment design is drawn from a Visual World eye tracking experiment we developed and conducted in Russian, English, and Spanish, with both adults and children of various ages, between 2018–2020 (Minor et al., 2022b). The aim of this study was a cross-linguistic comparison of three typologically different aspectual systems using the same picture stimuli and experiment design: in order to tease out subtle differences in the semantic representation of (in particular) the *perfective* verb forms in these systems, which are often grouped together under the same formal denotation, but are found to carve up narrative time in ways that are not easily captured by offline[5] judgments and truth conditions alone. The version of this study that we chose to replicate, namely the English, exemplifies a case where online processing data can help illuminate a muddled semantic landscape.

We contrasted the 'imperfective' English Past Progressive (e.g. *was baking, was painting*) with the 'perfective' Simple Past (e.g. *baked, painted*). The imperfective/perfective contrast is not binary in English to the degree that it is in, for example, Slavic languages (Gvozdanović, 2012); its grammatical rendering is somewhat lopsided, with the Past Progressive marking imperfective with an inflected *be*-AUX and a participial verb, and the Simple Past bearing only a tense suffix and no overt aspectual marker. The grammatical, or 'viewpoint', aspect of the Past Progressive is non-habitual continuous: it highlights the ongoing part of the event, and does not entail that the result state, or *telos*, of the event is ever reached (de Swart, 2012).

The interpretation of the Simple Past is less clear-cut: it is generally considered perfective (Van Hout, 2011), though stative verbs form an exception, and various semanticists recognize that the Simple Past does not always entail the culmination of an event (see e.g. van Hout, 2018; Martin et al., 2020; Martin & Demirdache, 2020). De Swart (1998) analyses the English Simple Past as aspectually neutral, with the (im)perfectivity of the verb being determined by its Aktionsart. When the Aktionsart is an accomplishment, however (as it is in this study), the Simple Past is interpreted as a perfective – after all, "culmination entailments are typically taken to be a diagnostic criterion for defining this aspectual class." (Martin & Demirdache, 2020).

This reading is supported by experimental work: Madden & Zwaan (2003), whose paper laid the foundation for the stimulus design of our study, found that the Simple Past constrained the mental representation of events. Magliano & Schleich (2000) found that the mental activation of events decayed faster if they were presented in the Simple Past form; and Bott & Hamm (2014)

[5] To avoid confusion, this paper will use the terms 'online' and 'offline' only to refer to behavioral measures collected in real-time, and after processing has taken place, respectively. Despite the widespread use of 'online' to mean 'on the Internet', we will refer to our webcam eye tracking study as being 'web-based', to mean conducted on the web/the Internet.

found that coercion of a (Simple) Past accomplishment predicate into an activity reading caused processing difficulty in English, but not in German.

However, there is also some evidence hinting that Simple Past accomplishment predicates do not *have* to be perfective: in a pragmatics study contrasting the telicity of particle verbs (e.g. *eat the apple up*) with that of corresponding simplex verbs (*eat the apple*), Jeschull (2007) found that adults' preference was at ceiling for a completion interpretation of particle verbs, but at chance for simplex verbs. In a study exploring the perfective interpretations of simple and complex verb forms describing change-of-state events in Hindi and English, Arunachalam & Kothari (2011) report that English speakers accepted partial-completion interpretations for Simple Past verbs approximately 50% of the time (patterning with the Hindi simple verb form, which does not entail event completion). In short: the jury is still out on the Simple Past, in both the theoretical and the experimental literature.

In order to better understand how the mental representation of accomplishment events is modulated by aspect in real-time, we chose to conduct a Visual World eye tracking study. We presented participants with two pictures of the same event: one in which the event is ongoing, and one where it has been completed. While viewing the pictures (the 'Visual World'), participants heard a sentence describing the event, in which the grammatical aspect of the verb was manipulated. Participants chose which picture best matches the sentence they heard (the offline result), and their approximate gaze fixations were measured throughout the trial. Two previous studies of this kind, Zhou et al. (2014) (Mandarin) and Minor et al. (2022a) (Russian), found that participants reacted to aspectual morphemes by looking towards their corresponding pictures immediately after hearing the morpheme, without waiting to hear all of the verb's arguments. In the case of Russian, this looking preference became statistically significant even before verb offset. This method therefore allowed us to see whether a more complex picture of online processing hides behind the varying offline judgments of the perfectivity of the Simple Past.

Our reasons for replicating this particular study were primarily practical: we were able to use the same stimuli and adhere to the original experimental design and data analysis as closely as possible, and conducting it in English allowed easy recruitment from a pool of over 40,000 eligible participants via Prolific.ac. However, the results of this study also made it an attractive candidate for replication: there was a stark and statistically significant difference between the two aspectual conditions, and there was a somewhat unexpected and intriguing *lack* of an effect of aspect in the Simple Past condition, with no detectable preference for either event type in both the online and offline results. Given that all the events were accomplishments, that the Simple Past is commonly analysed as a perfective, and that the design of the experiment, if anything,

encouraged participants to interpret the Simple Past in complementary distribution to the Past Progressive, we were surprised by the aspectual 'neutrality' (or ambivalence) of the Simple Past, and were interested in replicating this finding for its own sake.[6]

Our aim was to assess whether webcam eye tracking data performs well enough, in terms of temporal and spatial accuracy, to serve as a viable alternative to infrared eye tracking for Visual World experiments. In addition, we wanted to build our replication study using 'out-of-box' and open-source software tools with minimal customization, as a proof of concept that this kind of 'do it yourself' eye tracking experiment can realistically be built and run by anyone, without extensive programming experience or access to commercial platforms.

In the rest of the paper, we present a detailed Method of both studies, taking the original study as a default and specifying any adaptations made for the web-based version as and where appropriate. We then present the results, followed by a discussion of methodological factors affecting the temporal and spatial resolution of webcam eye tracking data, as well as participant retention and data quality.

2 Methods

All data, analysis scripts, materials, and experiment software are available at the Open Science Framework (https://osf.io/m395q/).

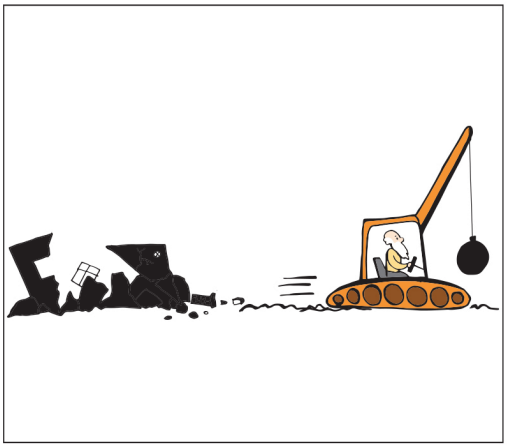
2.1 Materials and design

The experiment included 24 test trials and 24 filler trials, all consisting of a visual display and an audio stimulus. In the test trials, two pictures were presented side-by-side on a screen, representing two stages of the same event: one where the event is ongoing (cf. **Figure 1(a)**, and one in which it is completed (cf. **Figure 1(b)**).

The audio stimuli included a preamble and a target sentence, and were recorded by a female native speaker of British English in a sound-proof booth. The preamble was a short sentence in the past tense meant to set up a narrative context for the subsequent target sentence (e.g. *It was a crisp winter morning, There were many people shopping in town*, etc.). The target sentence was a

[6] Although it is arguably risky to try to replicate a null result using a novel, 'noisier' method, the unambiguously positive result in the Past Progressive condition provided us with a clear benchmark. We also had clear positive results in the perfective condition of our experiments in Russian and Spanish; as well as in a follow-up study contrasting the Past Progressive with the Past Perfect. This reassured us that the null result in the Simple Past condition reflected an absence of (strong) perfectivity, rather than e.g. the Ongoing Event picture generally being a better representation of the event than the Completed Event picture.

(a) Ongoing Event



(b) Completed Event

Figure 1: Visual display: 'An old man demolishing a house'.

transitive clause containing a subject NP (*grandma, grandpa, a girl* or *a boy*), a past-tense verb and an object NP consisting of an adjective and a noun (e.g. *a pretty flower, a new shirt*, etc.). The experimental manipulation was grammatical aspect, and as such there were two versions of each target sentence: one in which the verb had the Past Progressive form, and one in which the verb had the Simple Past form, cf. example 1.

(1) a. Grandpa was demolishing an old house.
b. Grandma demolished an old house.

All items in the experiment involved telic accomplishments (cf. Vendler, 1967; Dowty, 1979), i.e. events with a process stage and a well-defined result stage. The design of the experiment relies on the assumption that these two stages can be mentally represented as 'snapshots', and are quickly and easily recognized as such by participants.

In the filler items, the visual display featured two pictures of *different* events (e.g. *Grandpa chopped down a tree* vs *Grandpa blew out a big candle*), but with the same subject, in various combinations of ongoing event versus completed event. The preamble and target sentences in the fillers were similar to those in the test items. To counteract the experimental bias that would arise if hearing the auxiliary *be* would always uniquely predict the ongoing event picture, half of the filler items included a construction with the auxiliary *be* that described a completed event (e.g. *Grandma was successful in cracking open the nut, The boy was done with taking apart the wooden stool*); the other half of the fillers included a past tense verb describing the initial or intermediate

stages of an event (e.g. *The girl began to drink a glass of milk*, *Grandpa occupied himself in the strawberry patch*).

Two lists were created such that each test item appeared once in each list: in the Past Progressive form in one list, and the Simple Past form in the other. The position of the ongoing/completed event picture and the target picture was balanced across lists, and participants were randomly assigned to one of the two lists. Each list began with a filler item in order to acclimate the participants to the procedure; there were no practice items.

2.1.1 Web-based replication

The materials and design of the web-based replication were kept identical where possible, and minimally adapted to the constraints of the browser environment where necessary. To ensure that the pictures would have the same relative size and placement on the screen regardless of the display size,[7] the browser window was divided into a grid with 12 vertical columns (using the Bootstrap CSS framework), wherein each picture was centered in a container 5 columns wide, with a 2-column neutral space between them. The height and width in pixels of each container was recorded in the data output.

In the original infrared eye tracking study, the trial order was fixed to avoid any clustering of trials that could create a habituation effect; never more than two consecutive trials with the same target event type, or target picture presentation side, as well as alternating test and filler items. Because we expected the web-based replication to have a longer duration, and so possibly a stronger effect of boredom or habituation amid the overall higher level of noise, we pseudo-randomized the trials by ordering them in a list such that they formed blocks of four items that met all the balancing criteria we used for the original study. The order between these 8 blocks (but not within blocks) was randomized by participant.

The experiment was programmed using jsPsych (De Leeuw, 2015), which debuted its WebGazer-based eye tracking functionality with the release of version 6.3, in February 2021. The jsPsych framework organizes the various parts and functionalities of a behavioural experiment into modular scripts, or ‘plugins’; the eye tracking functionality is designed as an ‘extension’ that can be added to and run in the background of any other plugin. Additionally, the package includes

[7] The disadvantage of this approach is that the absolute size of the pictures will vary between participants, a noise factor that is all the harder to control for because display size cannot be automatically recorded through the browser (and screen resolution, which can, is not correlated with display size). Short of asking participants to measure the diagonal of their screen with a tape measure, we cannot know. However, giving the pictures the same absolute size and placement would likely cause them to be displayed incompletely or incorrectly on some devices. See section 3.2.2 for further discussion.

a plugin that initializes the webcam and locates the face and eyes in the center of the video feed; a calibration plugin, which trains a regression model to predict gaze location based on eye position; and a validation plugin, with which the accuracy of the prediction model is assessed. jsPsych version 6.3.1, released in April, also includes a forked version of WebGazer which was adapted to improve temporal resolution (see section 4.1.3 for a more in-depth discussion). Wherever possible, we used jsPsych's plugins and API 'out-of-box'; we lightly adapted the `audio-button-response` plugin to program the trials themselves, but were otherwise able to use the tools provided by jsPsych without customization.

The study was hosted on a JATOS[8] server owned by UiT – the Arctic university of Norway.

2.2 Procedure

2.2.1 Original infrared eye tracking study

Instructions were delivered verbally by the experimenter. Participants were calibrated once at the start of the experiment, using a 9-point calibration grid, which was validated by fixating a randomly presented succession of points on that grid. Each trial had a preamble phase and a target phase. During the preamble phase participants were shown a picture of a smiley face at the center of the screen and heard the preamble sentence. After that the trial proceeded to the target phase where two pictures were presented side by side on the screen. After a 500 ms preview the participants heard the target sentence, and chose one of the pictures by raising the corresponding hand (left or right). Participants' eye movements were recorded using an SMI RED500 eye tracker with an integrated 22-inch monitor, at a sampling rate of 120 Hz; offline responses were recorded manually by the experimenter. The experiment lasted approximately 6 minutes.

2.2.2 Web-based replication

After completing the demographic survey, granting permission to access the webcam and passing a browser and equipment check, participants were encouraged to ensure that they would be undisturbed for at least 15 minutes; their face would be brightly and evenly lit; and that they were sitting comfortably (see **Figure 2**).[9] They were not instructed to sit at a particular distance from their screen (though see section 3.2.2 for a possible approach to managing this in future work.)

---

[8] JATOS stands for "Just Another Tool for Online Studies" (Lange et al., 2015). It is a free, open-source backend tool for hosting and managing web-based studies.

[9] The instructional images for head positioning, posture, and lighting were taken from Semmelmann & Weigelt (2018) (https://osf.io/jmz79/).

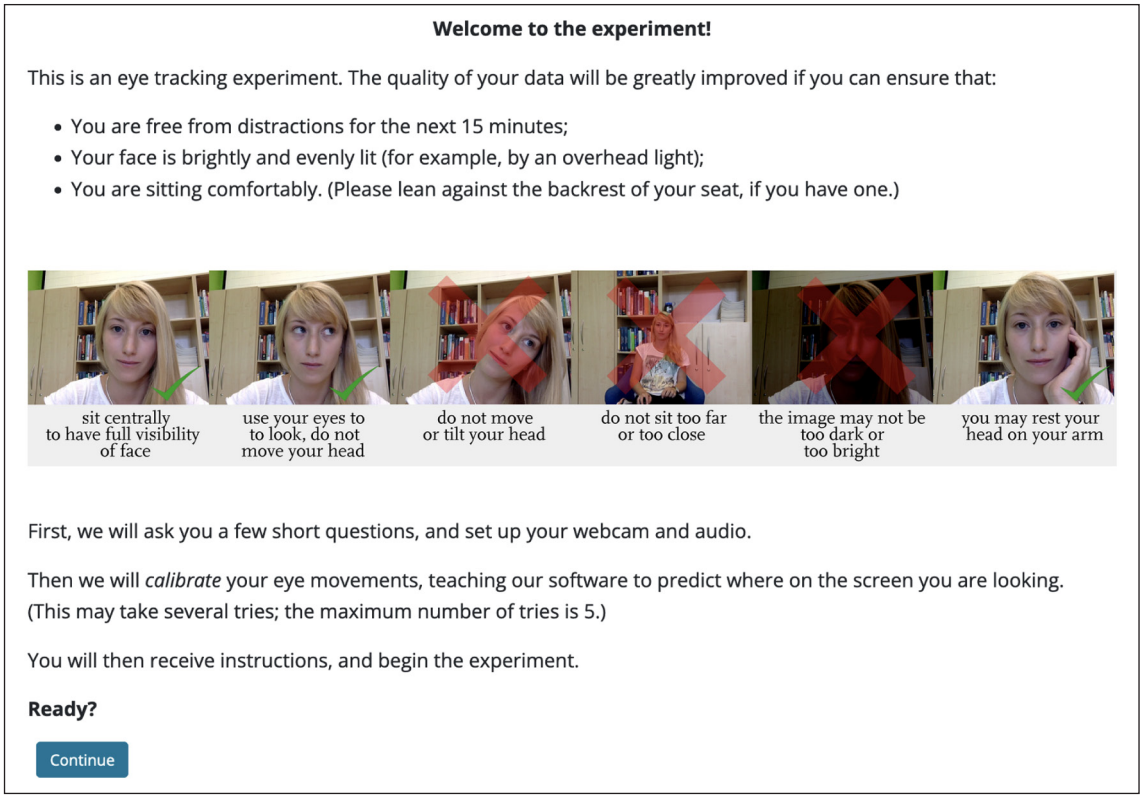


Figure 2: Instructions for posture and head positioning (the first page in the study).

Participants were then directed through a calibration phase. 15 points, 30px wide, were presented consecutively and in random order across the entire screen (between 10–90% of the screen dimensions). Participants were instructed to look at and click on each point; when clicked, the point would vanish and the next would appear. The validation of this calibration phase consisted of two consecutively presented points, 30px wide, centered approximately where the trial pictures would later be located on the screen. Each point was visible for 3 seconds: participants were instructed to simply look at the points without moving their head. Gaze predictions were generated starting 500 ms after each point appeared, to allow the eyes to saccade. For the participant to 'pass' the calibration, $>50\%$ of gaze predictions for each point had to fall within a 200px tolerance radius of that point.

Following the validation, participants received visual feedback on their performance: the tolerance radius around each point was made visible, and the raw gaze data of the validation was plotted onto the screen as green ('hit') and red ('miss') dots.[10] If the 50% threshold was

[10] In the jsPsych documentation this is recommended as a testing and debugging tool, but we chose to keep it in hope that participants would be able to interpret the feedback and adjust accordingly (e.g., if a participant's face is brightly lit on one side because they're sat beside a window, the accuracy on that side of the screen is likely to be much lower).

reached for both validation points, the participant could proceed to the experiment; if not, they were looped back to the start of the calibration phase. In addition, the sampling rate (i.e. the rate at which WebGazer generates gaze predictions) had to be at least 5 samples per second. If calibration was not successful within 5 attempts, the study was aborted.

This calibration procedure was repeated another three times throughout the experiment (once every 12 trials) to compensate for small head movements and resultant decay of the accuracy of the gaze prediction model (see section 3.2 for further discussion).

Following calibration, the experiment proceeded exactly as it had in the original study, with the exception that participants had to interact with the web page to advance to the next trial. The preamble phase of each trial began with a green fixation point at center screen, which had to be clicked to play the preamble audio. When the audio finished playing, the trial advanced automatically to the two-picture display, the target sentence started playing, and the participants' cursor was hidden. When the target audio ended, the cursor reappeared, and the participant had to select one of the pictures by clicking on it, which triggered the start of the next trial. Participants' eye movements during the target phase of the trial, and their picture choice, were recorded. The study duration was 15.32 minutes on average.

2.3 Participants

2.3.1 Original infrared eye tracking study

35 adult monolingual English speakers were tested in Edinburgh (Scotland), in December 2019. A further 31 adult monolingual English speakers were recruited and tested in Norway (Trondheim and Tromsø) in September 2020, giving a total of 66 participants. All the participants tested in Norway had spent less than 5 years in Norway prior to the experiment, and attested to having only elementary conversational proficiency in Norwegian.

All participants had normal or corrected-to-normal vision. Written consent was obtained from all the participants prior to testing; as compensation, the participants tested in Edinburgh received £5, and the participants tested in Norway received a cinema voucher or a gift card worth 120 Norwegian kroner (~$13.50).

2.3.2 Web-based replication

124 adult monolingual English speakers were recruited via Prolific.ac. The sample size was determined on the basis of the results of our pilot studies.[11] Several filters were applied on Prolific,

Feedback also bolsters intrinsic motivation and improves performance (see e.g. Dow et al., 2012): if participants can see that they have a healthy number of gaze samples, of which *just* under half appear to be within either tolerance radius, they will hopefully be motivated to keep going rather than discouraged because they failed to calibrate.

[11] Before running the replication study, we ran several smaller pilot studies in batches of 60 participants (the infrared study's sample size) at a time. By merging the results of two pilots, we found that we had reached a similar standard

to restrict who could access the study: participants had to be English speaking monolinguals who had spent most of their time before turning 18 in the United Kingdom; and they could only participate with a desktop computer (as opposed to a tablet or phone[12]) and a webcam.

Consent was obtained electronically by clicking a button labeled "I agree and Start" at the bottom of a reloadable information and consent page. Participants were paid £4 for an estimated study duration of 20 minutes (the actual study duration was ~15 minutes) if they successfully completed the entire study; they were paid £2 if they completed part of the study after a successful initial calibration, but were barred from finishing it after a failed recalibration. If they were unable to calibrate and start the study, they received no compensation. Participants were informed of this conditional payment structure in the information and consent letter. In total, 197 people started the experiment; of these, 39 (19.8%) dropped out before calibrating, usually because their browser, webcam, or audio output did not work. 16 (8.1%) dropped out after failing their initial calibration, 8 (4%) dropped out after failing a recalibration, and 124 (62.9%) successfully completed the experiment. The remainder refreshed the web page during the experiment, which blocked them from further participation.

2.4 Trial exclusions and data preparation

We inspected participants' accuracy in their picture choices in the filler trials to determine whether they merited exclusion. All 124 participants were >85% accurate, and so none were excluded. As in the original study, we excluded trials with >50% track loss (infrared version: 2 trials, 0.13% data loss): in the replication, this meant trials with >50% gaze predictions located outside the participant's screen dimensions (47 trials, 1.6% data loss).

2.5 Analysis

We coded the selection of the Ongoing Event picture in the Progressive condition and the Completed Event picture in the Simple Past condition as 'target', and the opposite choice as 'competitor'. To test whether the proportion of 'target' picture selections was significantly above chance in either condition, we fit two mixed effects logistic regressions (using the R package lme4(Bates et al., 2014; R Core Team, 2019) estimating the log-odds of a target response in the Past Progressive and Simple Past trials, with random intercepts for participants and items.[13]

To identify the time windows in which the probability of fixating on the target picture was significantly above chance, we performed a cluster-based permutation analysis for each condition

error of the mean as in the infrared results (though still a weaker overall effect): once we switched to the new version of WebGazer and found a much clearer gaze pattern with as little as $n = 24$, we decided to set the sample size at $n = 120$.

[12] We leave the question of whether this type of study could also be taken via tablet or phone to future research.

[13] We did not fit maximally structured random effects in our models because some failed to converge with both random intercepts and slopes (singular fit), so we applied only random intercepts across all models.

(see e.g. Huang & Snedeker, 2020; Yang et al., 2020). One advantage of this analysis over the more common growth curve analysis is that it gives an estimate of the time window (the titular 'cluster' of time bins) in which an effect is significant, without the researcher pre-defining which time windows to analyze (which can seriously affect the statistical outcomes; see e.g. Peelle & Van Engen, 2021; Huang & Snedeker, 2020).

We selected the data starting from verb onset to 2000 ms after verb onset; binned the data into 50 ms time bins; calculated the proportion of fixations on the target picture in each time bin; and then binarized that data by rounding up to 1 or down to 0. Next, we fit a mixed effects logistic regression for each time bin, to estimate the log-odds of fixations on the target picture. Items and participants were included as random intercepts, and an intercept term was included to represent the difference between the log-odd of fixations on the target picture and 0, which corresponds to chance (0.5) probability. Next, we clustered together consecutive time bins where the probability of fixating on the target picture was significant at $\alpha = 0.08$, on the assumption that these all exhibit the same effect;[14] and summed up their z-values to create a sum statistic for each cluster. Finally, we estimated how likely these clusters would be to occur by chance, under the null hypothesis that the probability of fixating on the target versus the competitor picture was at chance. We did this by creating a permutation distribution, whereby we randomly permuted the picture labels (target vs competitor) by participant and then repeated the regression and the clustering steps. This procedure was repeated 1000 times, yielding a distribution of sum statistics against which the statistics of the original clusters were compared. Clusters with $p < 0.05$ were considered significant.

2.6 Results

Table 1 shows the offline responses in both the original infrared eye tracking study and the WebGazer replication. In both studies, the preference for the 'target' Ongoing Event picture in the Past Progressive condition was almost at-ceiling, but the preference for either picture in the Simple Past condition hovers around chance level. The log-odds of selecting the target picture were significantly higher than 0 in the Past Progressive condition in both the original study (intercept $B = 6.24$, $SE = 0.77$, $Z = 8.09$, $p < 0.001$) and the replication (intercept $B = 5.25$, $SE = 0.59$, $Z = 8.86$, $p < 0.001$). In the Simple Past condition, the log-odds were not significant

[14] We chose to set the threshold at 0.08, rather than the customary 0.05, because relaxing the criteria for identifying clusters in the initial stage of the permutation analysis helps to find larger contiguous clusters. When we ran this analysis with a 'traditional' alpha of 0.05, we found several clusters grouped closely together, separated only by 1 or 2 time bins where an effect was not found: we did not find it plausible that an effect of aspect in looking preference (well after verb offset) should blink in and out of existence. With an alpha of 0.08, these clusters merged into one. The risk of false positives is sidestepped by the re-sampling portion of the analysis: the initial, "real" sum statistic is tested against a distribution of 1000 permuted sum statistics that were *also* found with alpha = 0.08, so this final significance test is no less strict.

in the original study (intercept $B = 0.26$, $SE = 0.3$, $Z = 0.86$, $p = 0.39$); and although the proportion of selections of the competitor picture in this condition increased by 10 percentage points in the replication, the log-odds of selecting the target picture still did not significantly deviate from 0 (intercept $B = –0.46$, $SE = 0.32$, $Z = –1.41$, $p = 0.16$).

	Infrared		**WebGazer**	
Event type	Prog	SPast	Prog	SPast
Ongoing Event	95%	46%	98%	56%
Completed Event	5%	54%	2%	44%

Table 1: Offline responses in the original study and the replication.

Figure 3 presents the online results from both the original study (a, b) and the WebGazer replication (c, d), starting from, and ending 2000 ms after, lexical verb onset. The dashed vertical lines mark average lexical verb offset, and the shading in (a) and (c) represents the time windows in which the probability of looking towards the target picture was significantly above chance. In these graphs, looks that fell outside either picture were filtered out: in the original study, that constituted 6.36% of gaze data, but in the replication, it was 27.95% (see section 4.2 for discussion).

Figure 3: Proportion of looks to the target picture in the Progressive condition ((a) and (c)), and in the Simple Past condition ((b) and (d)). Data in **(a)** and **(b)** were collected with an infrared eye tracker, data in **(c)** and **(d)** using WebGazer. The colored ribbons around the graph lines represent the standard error of the mean. Grey shading represents the time bins where probability of looks to the target picture was significantly above chance. The dashed vertical lines mark average lexical verb offset (559 ms in the Simple Past condition, 674 ms in the Past Progressive condition).

We used a cluster-based permutation analysis ($\alpha = 0.08$) to identify clusters of 50 ms time bins where the probability of fixating on the target picture was significantly above chance. In the original study, this analysis revealed one cluster from 500 to 2000 ms after lexical verb onset (sum $Z = 103.57$, $p < 0.001$) in the Past Progressive condition; and no clusters in the Simple Past condition. In the replication study, the analysis identified one cluster from 550 to 2000 ms (sum $Z = 133.8$, $p < 0.001$) in the Past Progressive condition, and no clusters in the Simple Past condition.

3 Discussion

We replicated a Visual World eye tracking study using browser-based experiment software and webcam eye tracking tools, and remote participants. We were able to fully replicate the results of the original study, including the approximate onset of the time window in which the probability of fixating on the target picture was significant (which was one time bin, or 50 ms, later in the replication). This is a marked improvement on the outcomes of earlier WebGazer replications of eye tracking tasks. In the following sections, we will try to contextualize and account for this improvement, discussing technical factors (particularly spatial and temporal resolution) on the one hand, and methodological factors affecting participant retention and overall data quality on the other.

3.1 Temporal resolution

Moreso than large spatial offsets (of which the Visual World Paradigm is much more forgiving than, say, eye tracking during reading), the primary concern with webcam eye tracking has been its low sampling rates and variable inter-sampling intervals – in other words, its poor temporal resolution. Semmelmann & Weigelt (2018) noted a higher temporal error when data was collected remotely, on participants' own laptops and browsers where processing load and hardware performance could not be controlled for. In their WebGazer replications of Visual World studies, Slim & Hartsuiker (2021a) and Degen et al. (2021) found 300–700 ms delays in the onset of the replicated effect, which they reasonably concluded would have to disqualify this eye tracking technique for use in any time-sensitive experiments.

There are methodological caveats to each of these studies that may, to some extent, account for their sluggish effect onsets – Semmelmann & Weigelt (2018) and Calabrich et al. (2021) recalibrated multiple times but only analysed the data of 28 and 14 participants respectively, Slim & Hartsuiker (2021) and Degen et al. (2021) had large datasets but no recalibrations, and so on. However, we think (and several of the aforementioned authors have indeed also speculated) that the biggest source of temporal noise in the data of these studies may have been courtesy of WebGazer itself. In their replication of a decision-making task with eye tracking (Krajbich et al., 2010), Yang & Krajbich (2021) found that as processing demands on the participant's browser and hardware increased over the course of the experiment, the time interval between gaze predictions increased dramatically, peaking at 972 ms ($SD = 107$ ms). They made an adjustment to the

WebGazer software itself, whereby the process that generated gaze predictions was decoupled from the main process that updated with every new animation frame – a process that is highly vulnerable to timing delays when the browser has to juggle several intensive tasks. With this adjustment, they were able to achieve much higher and more stable sampling rates. Shortly after Yang & Krajbich (2021)'s results became available, an update to jsPsych was released (version 6.3.1, April 10 2021) which included a forked and modified version of WebGazer: like Yang & Krajbich (2021), the developers had found that the method WebGazer relies on to generate gaze predictions created a processing bottleneck that caused serious temporal errors, and adjusted the code to resolve the problem.[15]

As we had been running pilots of this study using custom eye tracking plugins written in the jsPsych framework, we were able to switch to jsPsych v6.3.1 immediately after it came out. Like the other replication studies, the experimental effects in those pilots resembled those of the original, but slower, weaker, and noisier. The temporal resolution of our data and the onset of our experimental effect improved dramatically as a result of implementing the experiment in jsPsych v6.3.1[16]: due to several minor methodological improvements, but mostly, we expect, due to the modified WebGazer code.

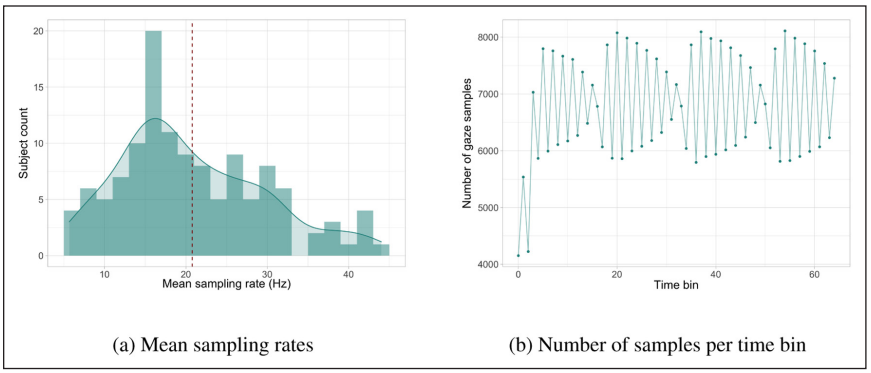


Figure 4: Webcam study gaze sampling rates. a) Histogram and density plot of participants' mean sampling rates. The red vertical line represents the grand average sampling rate: 20.73 Hz (SD = 8.99). b) Total number of gaze samples per time bin. Due to participants' varying (but consistent) sampling rates, the number of gaze samples oscillates between time bins.

[15] Josh de Leeuw, the main developer of jsPsych, clarified on 15/6/2021 (discussion #1892 on the jspsych/jsPsych github forum): "As far as performance goes, I think we [Yang and de Leeuw] both applied similar modifications to webgazer. […] The major change we both made to our respective forks is that we disabled webgazer's automatic loop so that webgazer is no longer trying to provide an updated prediction with every animation frame, and instead we just invoke webgazer's prediction algorithm at a regular interval. This seems to actually speed up the rate at which calculations can be done. And, perhaps even more importantly, using requestAnimationFrame was causing blocking in jsPsych's timing, so if a participant had a particularly poor computer – or even a good one – the timing of experiments could become really bad really quickly (see issue #1700)."

[16] A note for users of Gorilla.sc: Will Webster, a software developer at Gorilla/Cauldron Science, confirmed that his team is aware of this issue and is working on forking, modifying, and integrating WebGazer into Gorilla's own timing system (6/8/2021, personal correspondence).

By filtering out participants with a very low sampling rate (< 5 Hz), and relying on jsPsych's version of WebGazer, our participants had an average sampling rate of 20.73 Hz (*SD* = 8.99): about one gaze prediction per 48 ms. Though, as can be seen in **Figure 4**, the spread of our participants' sampling rates spans 5 to 45 Hz; and as a result, the number of data points per 50 ms time bin oscillates by as many as 2000 samples. (This oscillation could conceivably be why the significant effect window in **Figure 3(c)**, as identified by the cluster-based permutation analysis, starts 50 ms later compared to the infrared study!)

In webcam eye tracking, the sampling rate is effectively limited by the frames-per-second (fps) rate of the webcam – that is, WebGazer *can* generate predictions at a higher rate, but they may not reflect a 'real' observation of the eyes. Most consumer-grade webcams sample at 15 to 30 fps (though more expensive ones can go up to 60 fps); the real-time sampling rate is, however, affected by the processing load of the participant's device, so an actual fps and WebGazer sampling rate of 60 Hz is unlikely to occur. Semmelmann & Weigelt (2018) reported mean sampling rates of 18.71 fps (*SD* = 1.44) in their in-lab dataset, and 14.04 fps (*SD* = 6.68) in their remotely collected dataset; Yang & Krajbich (2021), using their modified version of WebGazer, report an average of 24.85 ms between gaze predictions (*SD* = 12.08), which converts to a 40.2 Hz sampling rate.

Given that infrared eye tracking systems are usually sampling at anywhere between 100–500 Hz, is this sampling rate sufficient? For a Visual World study, where the measure of interest is usually the proportion of *fixations* in a particular time window rather than saccadic eye movements, it appears that it is. The added value of very high sampling rates might even be doubtful in this paradigm: a fixation lasts 100–300 ms on average and even an express saccade will take at least 80 ms to launch, so how high a temporal resolution is really necessary? For instance, Dalmaijer (2014) conducted a method study with the EyeTribe eye tracker, which sampled at max. 60 Hz. He concluded that a 60 Hz sampling rate was good enough for research questions centered around fixation data. Ouzts & Duchowski (2012) compared two eye tracking datasets with different sampling rates, and recommend downsampling to the lower rate rather than upsampling to the higher rate (as is common practice); higher does not always equal better, especially if it means padding the data by splicing data points. Andersson et al., (2010) simulated various sampling rates in experiments with varying demands for temporal precision, and show that error resulting from lower sampling rates can be mitigated with higher power.

3.1.1 Temporal precision

Beyond WebGazer's temporal resolution, there is the timing precision of the experiment itself to consider.

The *accuracy* of the software's timing performance is not a grave concern for studies where the measure of interest is the difference between two or more conditions, such as the one presented here: any constant timing offset, or lag (usually arising from hardware characteristics) is canceled out. The *precision*, or variable error of the timing offset, does need to be accounted for. de Leeuw & Motz (2016), comparing participant response times in a Javascript versus a Psychophysics Toolbox application, note that where there is a noticeable difference, it is mostly an increase in lag that's relatively consistent across trials, and not an increase in variability between trials. In other words: poorer accuracy, but not poorer precision. Likewise, Slote & Strand (2016) found that variation in the measurement error of audio stimulus onset in a Javascript experiment could be limited to less than 5 ms, even when processing load was high, by using WebAudio API to schedule audio presentation. Bridges et al. (2020) compared the timing performance of several popular behavioral science software packages, both in-lab and web-based, on a range of operating systems and browsers. They found that jsPsych showed an inter-trial variability of precision in the range of 3.2–8.4 ms in all browser/operating system configurations. More generally, they note that a problem which seems to affect all online software packages to various degrees is the exact synchronization of audio and visual stimuli, a task for which Javascript is not ideal (cf. Anwyl-Irvine et al., 2021a for additional data).

In future work, we may be able to use the participant sound card's own estimation of the audio output latency to better understand the temporal accuracy and precision of this experimental set-up, but with a within-subjects experimental design and a statistical model taking random participant effects into account, timing offsets and variation of this size should not hinder a clear interpretation of the data.

3.2 Spatial resolution

The other question to address in evaluating the performance of this web-based method is how accurately it captures gaze location. Though our results indicate that WebGazer's spatial resolution is good enough to capture the expected effect in a two-picture paradigm, the fact that 28% of the replication data were looks outside either of the pictures (vs 6.3% in the original study) indicates that WebGazer remains a blunter instrument than infrared. This requires a balancing act: in order to minimize the risk of looks being misclassified, we had separated the two pictures by 20% of screen width, with the result that the majority of non-picture looks concentrated in this area (see **Figure 5(b)**).

It could be (though this is pure speculation) that, because our participants' average screen size was much smaller than the display of our infrared eye tracker, they used more of their peripheral vision (which can and does contribute to object recognition and scene perception; see e.g. (Rosenholtz, 2016)) to perceive the pictures, resulting in more looks at center screen. With

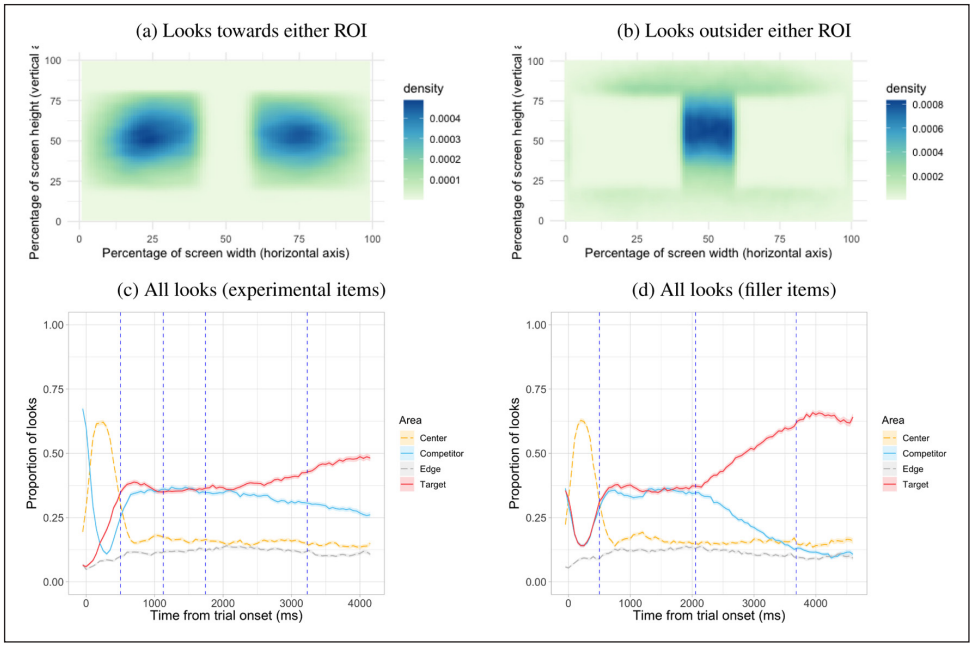


Figure 5: Density plots for **(a)** looks towards either Region Of Interest, and **(b)** looks outside either ROI (webcam study). Gaze and picture placement coordinates were computed as a percentage of the screen width and height. Two participants (#44 and #95) have been excluded from these graphs because their relative picture placements were not aligned with the others, possibly because they exited fullscreen mode and adjusted their browser window dimensions. Graphs **(c)** and **(d)** plot looks towards the target picture (red), the competitor picture (blue), the center column of the screen (yellow), and the remaining edges of the screen. The vertical dotted lines mark, from left to right: audio onset, verb onset, verb offset (in (c) only), and audio offset.

Figures 5(c) and **5(d)**, we can rule out that the high density of looks towards the center is driven by the reappearance of the cursor at audio offset: the proportion of looks to the center is high at the start of the trial (as expected), then drops sharply as soon as the audio stimulus begins, and remains at just over 20%.[17]

What *is* evident is that spatial resolution is not equal everywhere: Semmelmann & Weigelt (2018), Yang & Krajbich (2021) and Slim & Hartsuiker (2021b) found that fixation targets near the corners of the screen had significantly higher gaze offsets, and Semmelmann & Weigelt (2018) report that gaze predictions for targets near the bottom of the screen have considerable

[17] A minor but interesting note on Figures 5(c) and 5(d) is that they show very different proportions of looks to the target vs. the competitor picture in the first ~200 ms of the trial. This could in part be due to spillover from the previous trial – experimental and filler items always alternate, and the preference for the target picture is much greater in the fillers. But as can be seen in Figure 4(b), the first 3 time bins of the trial have much fewer gaze samples than the rest – as if WebGazer has to 'warm up' in the first 150 ms before settling into a regular sampling pattern. Something worth keeping in mind while designing experiments with WebGazer.

offsets towards the top. This may be due to interference from eyelids or eyelashes, or simply a consequence of the positioning of the webcam (generally at the top of the screen).

Despite these caveats, replications of four-picture Visual World studies (Degen et al., 2021; Slim & Hartsuiker, 2021b) so far indicate that WebGazer's spatial resolution can accommodate a more crowded display – but this primarily depends on the quality of the calibration.

3.2.1 Calibration

In conventional infrared eye tracking experiments, eye movement is tracked by reflecting a near-infrared light beam off the eye, and measuring the distance between the resulting glint (a.k.a. first Purkinje image) and the center of the pupil. This method is so accurate that for most experiments, the tracker is calibrated only once, at the start; provided the participant does not move and ambient light conditions remain relatively stable, recalibrations are usually not necessary. In eye tracking using visible light spectrum cameras, however, gaze prediction is inevitably less accurate: WebGazer does it by isolating the webcam image of the eyes as detected by a facial features recognition algorithm, reducing it to a 120-pixel grayscale eye feature vector, and supplying that to the gaze prediction model. This approach is more vulnerable to variable or uneven lighting, small head movements, etc., and so the accuracy of the gaze prediction model can be expected to decay significantly over the course of the experiment – see e.g. Degen et al., 2021's webcam replication study, in which they did not recalibrate at any point during 54 trials. At the other extreme sit Semmelmann & Weigelt (2018), who recalibrated their participants before every block of trials, leading to about half of their study's duration ($M = 43.54$ minutes) being spent on calibrating; they noted this was a somewhat arbitrary choice which seemed to wear out their participants, and marked the issue of how often to recalibrate as an important one to answer in future work.

Rather than set a fixed number of recalibrations every n trials, Yang & Krajbich (2021) opted for conditional recalibrations: every 10 trials, their participants would see three validation dots (each visible for 2 seconds). If participants fell below the 'hit' threshold (70% of gaze predictions within 130px of the validation dot) for four dots in two validations, they would recalibrate. Analysing the 'hit' ratios of their validation trials, they found that the ratio dropped right after calibration, but declined very slowly at every successive validation trial.

In order to understand the rate of calibration accuracy decay in our own experiment design, and to determine the number of recalibrations needed for our replication experiment, we conducted a pilot wherein participants were calibrated only once at the start, and the calibration accuracy was measured with a validation after every second trial. **Figure 6** presents the results of that pilot: the accuracy drops off immediately after calibration, and continues to decay quite rapidly thereafter.

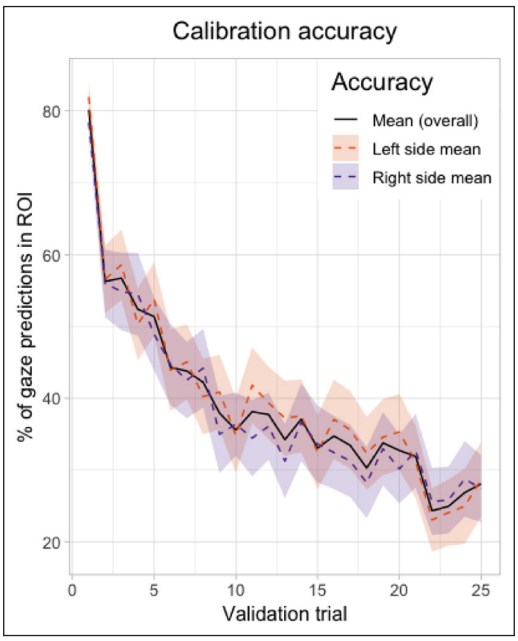


Figure 6: Decay of the accuracy of the initial calibration (webcam study). The shaded ribbons represent the standard error; in this pilot (as in the replication study), the presentation order of the validation points was not randomized, and left was always presented first.

Wanting to mitigate this decay, but also to avoid exhausting our participants with frequent recalibrations, we chose to recalibrate 3 times, or once every 12 trials. Because it appears that the rate of decay varies by experiment design, we would recommend piloting any webcam eye tracking experiment with a similar procedure, to determine the optimal number of recalibrations.

It is worth noting that Yang & Krajbich (2021)'s approach to tracking calibration accuracy decay was quite different from ours, and placed more performance demands on their participants: the consequence of too many validation 'misses' was a recalibration, and validation success or failure was communicated through colour (the validation point turning green for a 'hit', and red for a 'miss'). In our pilot, participants received no feedback on their performance during inter-trial validations, and experienced no consequences for slacking off. This could at least partially explain the steeper drop-off in calibration accuracy in our pilot. Having now established, through the collective effort of the various method studies cited here as well as our own, that *technologically* WebGazer can achieve the necessary spatial accuracy for Visual World studies, the development of *behavioural* best practices for calibration and validation during webcam eye tracking experiments would be a useful focus for future work.

3.2.2 Screen size and relative stimulus size

It is worth revisiting an experimental design flaw noted in footnote 6 (section 2.1.1): participants' screen size was not controlled for, and because we sized the screen contents relative to the

browser window dimensions (in the case of the picture stimuli) or in pixels (in the case of the calibration points), the absolute size of the screen contents also varied between participants. We cannot guess at how great this variation is: because of a bug in our experiment software, we unfortunately did not collect accurate information about whether participants used a laptop or a PC with an external monitor. In a subsequent, near-identical study (wherein this bug was fixed), we found that 95% of participants used a laptop, giving reasonable hope that the size variation is modest.

How to address this problem in future work? In the case of the calibration points: thus far, the jsPsych calibration and validation plugins only accept number of pixels as a measure of point size. (Likewise the size of the tolerance radius around validation points.) While that means that the absolute size of the points will differ depending on screen resolution, the question is whether sizing them by some measure other than pixels will help. The eye feature vector constructed by WebGazer to track gaze is, after all, also measured in pixels. However, it should be possible to adjust the plugins to allow the size of the points (and of the tolerance radius around them) to be computed as a percentage of screen size, or some other bespoke measure.

Stimulus size can be more easily standardized. When we ran this study, this solution was not yet available, but jsPsych has since introduced a plugin for the Virtual Chinrest (based on Li et al. 2020), which can be used to measure the distance between the participant and their screen, as well as standardize the the jsPsych page content to a known physical dimension. This could potentially be a good way to ensure all participants see pictures of the same absolute size, regardless of screen size. Though it does add another 'hoop' for participants to jump through, an issue we'll discuss in the next section.

3.3 Participant retention and data quality

One of the major selling points of webcam eye tracking, as previously stated, is that data collection could potentially be much quicker and more efficient than its lab-based counterpart. Researchers no longer need to invite participants to the lab, to be tested one-by-one, or travel to reach their target demographic; this can also save a lot of money. Given a large remote participant pool (such as Prolific, particularly for English speakers), data collection may be completed within 1–2 hours as opposed to weeks or months. However, several webcam eye tracking studies (e.g. Anwyl-Irvine et al., 2021b; Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021; and Slim & Hartsuiker, 2021b) remark on their experiments' high participant attrition rates as a cause for concern: 62% in Semmelmann & Weigelt (2018), 61% in Yang & Krajbich (2021), and 72% in Slim & Hartsuiker (2021b), which Anwyl-Irvine et al., (2021b) cited as motivating their development of MouseView.js.[18] Not only do high attrition rates undermine the time- and cost efficiency

[18] MouseView.js is a Javascript library which blurs the display to mimic peripheral vision, but lets participants use their mouse pointer to move a sharp, fovea-like aperture.

of browser-based eye tracking for the researcher, they also suggest that the experiment is too difficult, uncomfortable, and/or long for the average participant – a problem worth resolving because the remaining sample may be skewed ('survival bias'), but also for its own sake. For our experiment, we therefore sought to improve the experience of taking part on the participants' side, while also filtering out participants with sub-optimal equipment set-ups as early as possible in the study flow. Of our 197 participants who began the study, 73 (37% attrition) did not complete it; if we remove the 39 participants who were prevented from advancing to the initial calibration because of equipment issues, the attrition rate drops to 21.5%. Here we consider a number of factors that we believe impact participant retention and overall data quality.[19]

Webcam eye tracking demands a lot more from a participant than the average survey or even reaction time experiment: for best results, they are asked to rest the computer on a flat surface, assume a posture they can comfortably maintain without moving for several minutes, adjust lighting if necessary, and close processing-heavy apps running in the background. For participants recruited via platforms such as Amazon Mechanical Turk or Prolific, wasted time means lower earnings; with an all-or-nothing renumeration policy (e.g. Semmelmann & Weigelt (2018) only paid the participants that completed the entire study, namely $4 for an average study duration of 43.54 minutes), participants will quickly give up if they risk earning nothing after failing a recalibration. By paying our participants well above minimum wage, and by offering 50% payment if participants failed a recalibration, we hoped to convey appreciation for the concerted effort it takes to cooperate with the experiment design, and to incentivize that effort.

Beyond these pragmatic considerations, participants' tolerance for boredom may be lower. Many of the recent articles, webinars and blog posts reviewing methods and best practices of online research emphasize the limited 'patience time window' of participants: the consensus, insofar as there is one, seems to be roughly 20 minutes (cf. e.g. Kochari, 2019; and a recent webinar on web-based eye and mouse tracking by Gorilla.sc). In a survey of 103 Germans, Sauter et al. (2020) found that 44% would abort an online study paying minimum wage if it took longer than 15 minutes; this figure rose to 79% for 30 minutes. On the other hand, Jun et al. (2017) and Chandler & Kapelner (2013) find that if the study is considered interesting or meaningful that may mitigate effort, boredom, and fatigue. Without the performance pressure induced by a lab environment and direct supervision, the onus is on the researcher to design an experiment that is both short and pleasant to interact with.

In that regard, the amount of time and effort spent (re)calibrating is probably where there is most room for improvement. One perk of this process is that passing a calibration accuracy threshold amounts to a built-in gate-keeping mechanism: few bad faith participants will struggle through

[19] For a more general overview and cost-benefit analysis of conducting web-based behavioral studies, see e.g. Sauter et al. (2020), Eyal et al. (2021), and Gagné & Franzen (2021).

repeated calibrations only to deliberately ignore the instructions for the experimental trials. Nonetheless, Semmelmann & Weigelt (2018) notes that 'gamifying' the calibration procedures in webcam eye tracking experiments would do much to improve participants' enjoyment (and performance) – in fact, Xu et al. (2015) did just that: they developed two short video games based on well-known game formats. One game, based on Angry Birds, required a high degree of accuracy (the goal was to "train a powerful gaze-controlled gun" with which to take down the birds); the other, based on Whack-a-mole, had a more forgiving threshold for successful 'hits'. Xu et al. (2015) used these games to advertise two versions of the same picture classification task: one longer, more demanding task yielding high-accuracy gaze data, and one shorter, easier task that yielded cruder data, but which was also much more popular on Amazon Mechanical Turk and attracted and retained more participants. By combining both data sets and post-hoc data processing, the authors were able to obtain satisfactory results. xLabs, a now-defunct company that offered webcam eye tracking for marketing research, calibrated users by letting them click on animated crawling ants or floating balloons, and validates by visualising their real-time gaze predictions as a 'laser' with which to squash or pop them.[20] This kind of gamified (re-)calibration process would also make the experiment design more suitable for children.

Finally, our aim was to mitigate some of the increased noise that is inevitably inherent to remote webcam eye tracking. Where possible, we opted for the thriftier approach of rejecting participants with sub-optimal equipment set-ups before starting the experiment, rather than removing them post-hoc. Potential participants can be filtered by browser features such as browser type and version, screen resolution, display refresh rate, and support for essential software libraries such as WebAudio API before starting the experiment.[21] Furthermore, the `samples_per_sec` variable in the validation plugin's data output can be used to filter out participants with a low sampling rate, which we take to be a symptom of an unsuitable setup – whether due to aged hardware, high CPU load, sub-optimal combination of operating system and browser, or some other factor. We set our threshold at 5 samples per second, but e.g. Madsen et al. (2021) set it at 15. In experiments with an expected smaller effect size or that require a more fine-grained spatial or temporal resolution, filtering by a relatively high sampling rate may be a prerequisite for obtaining interpretable data.

With regards to data post-processing, eye movement classification and raw gaze data smoothing algorithms constitute their own subfield within eye tracking research, and a review of that literature lies outside the scope of this paper (but see e.g. Salvucci & Goldberg, 2000; Tafaj et al., 2012; and Hessels et al., 2017). However, we note that Xu et al., 2015 extracted fixations from their raw gaze data through meanshift clustering, i.e. algorithmically identifying

[20] Much of the code behind these calibration games is still available via the company's Github page.

[21] As of version 7.1, jsPsych has a dedicated plugin for this.

and assigning gaze data to spatio-temporal clusters and labeling the cluster center as one fixation. To evaluate this approach, they selected 1000 random pairs of images and participant gaze data from the infrared eye tracking dataset they were using for comparison subject/image pairs (Judd et al., 2009), and obtained 'ground truth' fixation locations on the images from the gaze data. They then permuted that data to resemble webcam eye tracking data by subsampling it to 30 Hz and adding position noise; extracted fixations using their meanshift algorithm; and compared the results to their ground truth fixations. The algorithm was able to estimate these fixations reasonably well, which suggests that it is worth considering as a noise-reduction tool for webcam eye tracking data going forward.

Although we did not process our data beyond the procedure followed in the original infrared study, we suggest that there are several ways to filter out participants with low-quality data during data preparation: for example, by plotting and inspecting the distribution of a certain performance metric, and discarding participants below a certain cut-off point. Madsen et al. (2021)[22] visualised the raw gaze data of their participants along the horizontal and the vertical axes, coding position as brightness and with time on the x-axis and subject on the y-axis; subjects were sorted top-to-bottom by their score on a comprehension test. In their plots, high-performing subjects clearly exhibited a stereotypical pattern of eye movements, which gradually fades out as performance drops. In a Visual World study such as the one we present here, with an equal or greater number of filler trials than experimental trials, performance on the fillers could serve as a similar heuristic. Though accuracy of participants' offline responses on the filler trials was at ceiling, and thus less effective as a metric, one could use a summary metric of the online filler data instead, e.g. the proportion of looks towards any Region of Interest.

4 Conclusion

We have presented a web-based replication of a Visual World eye tracking study, demonstrating that it is possible to obtain results that approach laboratory-grade effect sizes and onsets, using free, open source and beginner-friendly software tools. We have also shown how with a few methodological and experiment design adjustments, the overall user experience and success of such a study can be improved. We thereby add to the rapidly growing body of work investigating the possibilities and limitations of WebGazer and remote webcam eye tracking studies, which in a short time has led and will undoubtedly continue to lead to better code, experimental protocols, participant experiences, and research outcomes.

[22] See page 9 of their Supplementary Information.

Supplementary information and materials

Demographic data

Before the start of the web-based experiment, we administered a short demographic survey and recorded participants' browser type and version,[23] their operating system type and version, and their screen resolution. In order to limit the collection of personal data that has no well-motivated bearing on the research question, we chose not to record sex or gender. Nor did we record exact age, choosing instead to bin participants into 5 age groups: 18–30, 31–43, 44–56, 57–69, and 70+. The demographic data is given in **Table 2**, with the number of participants for each category given between parentheses.

Age	Vision	Browser	OS
18–30 (52)	Normal, uncorrected vision (67)	Chrome (113)	Windows (84)
31–43 (35)	Glasses (49)	Firefox (11)	MacOS (31)
44–56 (29)	Contact lenses (4)		Chrome OS (7)
57–69 (8)	Abnormal, uncorrected vision (4)		Linux (2)
70+ (0)			

Table 2: Number of participants per category for Age group, Vision, Browser type, and Operating System.

Correlation of sampling rate and calibration accuracy

Spearman rank correlation of mean calibration accuracy (for successful calibration trials only) and mean sampling rate by participant (see **Figure 7**). Because Slim & Hartsuiker (2021b) found a strong correlation between their participants' calibration scores and frames-per-second rate ($R = 0.852$, $p < 0.001$), we ran a similar correlation on our data. At $R = 0.16$, and $p = 0.074$, this correlation was not significant; but since Slim & Hartsuiker (2021b)'s minimum threshold for calibration success was 5% for one validation point, and ours was 50% for two points, this is not surprising.

[23] At the time this study was conducted, the only browser systems that it could reliably be conducted on were Chrome and Firefox. Likewise, Yang & Krajbich (2021) report that of their 49 participants, 45 used Chrome and 4 used Firefox. The software libraries that webcam eye tracking experiments rely on – particularly, WebGazer and WebAudio API – are now becoming available to a wider range of browsers, such as Safari, Edge, and Opera.

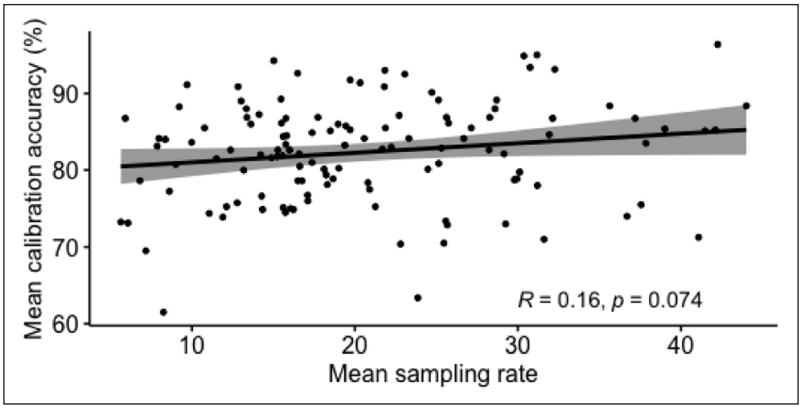


Figure 7: Spearman rank correlation of participants' mean sampling rate and mean calibration accuracy.

Sentence stimuli

The sentence stimuli used in both the original and the replication study: see **Table 3** for the experimental items, and **Table 4** for the filler items.

Preamble	Sentence
It was a crisp winter morning.	Grandpa was building a big snowman. Grandpa built a big snowman.
It was playtime at the school.	The boy was coloring a pretty picture. The boy colored a pretty picture.
It was time for lunch.	Grandma was slicing a juicy watermelon. Grandma sliced a juicy watermelon.
It was a bright sunny day.	Grandpa was digging a deep pit. Grandpa dug a deep pit.
It was playtime at the school.	The girl was drawing a slender vase. The girl drew a slender vase.
There were jobs to do around the house.	Grandpa was drilling a big hole. Grandpa drilled a big hole.
It was a holiday weekend.	Grandpa was fixing the old fridge. Grandpa fixed the old fridge.
It was a holiday weekend.	Grandma was hanging a beautiful painting. Grandma hung a beautiful painting.

(Contd.)

Preamble	Sentence
There was going to be a party.	Grandpa was ironing a clean shirt. Grandpa ironed a clean shirt.
It was a crisp winter morning.	Grandma was knitting a new jumper. Grandma knitted a new jumper.
It was a crisp winter morning.	Grandpa was lighting a cosy fire. Grandpa lit a cosy fire.
It was the middle of the afternoon.	Grandma was locking the side door. Grandma locked the side door.
The weather was nice and warm.	The girl was opening a big window. The girl opened a big window.
There were jobs to do around the house.	The girl was painting a high wall. The girl painted a high wall.
It was a bright sunny day.	Grandma was planting a pretty flower. Grandma planted a pretty flower.
It was the first period (at school).	The boy was sharpening a thin pencil. The boy sharpened a thin pencil.
There was going to be a party.	The boy was sweeping the narrow corridor. The boy swept the narrow corridor.
It was early in the morning.	The boy was cleaning the front room. The boy cleaned the front room.
The weather was nice and warm.	Grandma was watering a green bush. Grandma watered a green bush.
It was a rainy day outside.	Grandma was baking a lovely cake. Grandma baked a lovely cake.
It was a dark night with no hint of a breeze.	The boy was burying a wooden chest. The boy buried a wooden chest.
There were many people shopping in town.	The girl was buying a new phone. The girl bought a new phone.
It was time for lunch.	The girl was eating a tasty fish. The girl ate a tasty fish.
It was the middle of the day.	Grandpa was demolishing an old house. Grandpa demolished an old house.

Table 3: Sentence stimuli (experimental trials).

Preamble	Sentence
It was early in the morning.	Grandpa was satisfied that the candle was blown out.
It was the middle of the afternoon.	Grandma was successful in cracking open the nut.
There were jobs to do around the house.	The boy was done with taking apart the wooden stool.
It was a beautiful quiet evening.	Grandpa was unconcerned that the old bridge had been destroyed.
It was early in the morning.	The girl was happy with her newly cut out flower.
It was the middle of the afternoon.	The boy was pleased with his super tall tower.
There was going to be a party later.	Grandma was impressed with the beautiful dress she had sewn.
It was a rainy day outside.	Grandma was halfway through cutting the sleeves off the shirt.
The weather was nice and warm.	Grandpa was tired after chopping down the tree.
It was a bright and sunny day.	The girl was proud that she managed to swim across the river.
It was break time at the school.	The girl was finished with her glass of milk.
It was the middle of the afternoon.	The girl was ready to eat an orange.
The weather was nice and warm.	The boy enjoyed himself photographing nature.
It was a holiday weekend.	Grandpa concentrated on preparing dessert.
It was early in the morning.	The girl worked on cutting out a flower.
There were jobs to do around the house.	The boy wanted to take apart the old wooden stool.
It was the middle of the afternoon.	Grandpa relaxed and read a book.
The weather was nice and warm.	Grandpa occupied himself in the strawberry patch.
It was a beautiful quiet evening.	The boy planned to saw up the log for the fire.
It was the first period at school.	The boy got started on constructing a tower out of blocks.
It was playtime at the school.	The girl wanted to put together the pretty toy castle.

(Contd.)

Preamble	Sentence
It was late in the afternoon.	The girl decided to burn a blue notebook.
It was the middle of the day.	The girl began to drink a glass of milk.
It was a bright and sunny day.	The girl started to blow up the green balloon.

Table 4: Sentence stimuli (filler trials).

Data availability

All materials relating to this study, including the raw data files, the tidy, analysis-ready dataframe, the R script for the data analysis and visualisation, the stimuli, and the codebase for the experiment, are made available through the Open Science Framework (https://osf.io/m395q/).

Ethics and consent

All procedures performed in studies involving human participants were in accordance with the ethical standards of the Norwegian Centre for Research Data and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent was obtained from all individual participants involved in the study.

Acknowledgements

We gratefully acknowledge the work of Onur Ferhat, whose custom eye tracking plugins formed the basis of our pilot experiments. Thanks to Alain Schoorl for additional programming support; to the members of the Øyelab for their feedback and support throughout; to the editor and the reviewers for their helpful feedback; and to the hundreds of participants who contributed to these results.

Thanks also to the Linguistics departments at NTNU in Trondheim, and at the University of Edinburgh, for hosting us during our infrared eye tracking data collection trips. During our stay in Edinburgh, in December 2019, the academic staff were striking in protest against inequitable pay, extreme workloads, casualisation, and pension cuts: we stand in solidarity with their ongoing fight for a fairer and better workplace.

Funding information

The authors were funded by the Research Council of Norway FRIPRO project 275490 *Modal Concepts and Compositionality: New Directions in Experimental Semantics*.

This research was also supported through funding from the BLINK project (Marie Skłodowska-Curie Actions Individual Fellowship 2018-2021).

Competing interests

The authors have no competing interests to declare.

Author contributions

Conceptualization, M.V., S.M., and G.R.; methodology, M.V., S.M., and G.R.; software, M.V.; formal analysis, S.M.; visualization, S.M and M.V.; investigation, M.V.; data curation, M.V., S.M.; project administration, M.V.; validation, S.M.; writing – original draft, M.V.; writing – review and editing, M.V. and G.R.; supervision and funding acquisition, G.R.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, *38*(4), 419–439. DOI: https://doi.org/10.1006/jmla.1997.2558

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. DOI: 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). DOI: https://doi.org/10.16910/jemr.3.3.6

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, Jo. K. (2021a). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior research methods*, *53*(4), 1407–1425. DOI: https://doi.org/10.3758/s13428-020-01501-5

Anwyl-Irvine, A. L., Armstrong, T., & Dalmaijer, E. S. (2021b). Mouseview. js: Reliable and valid attention tracking in web-based experiments using a cursor-directed aperture. *Behavior research methods* (pp. 1–25). DOI: https://doi.org/10.3758/s13428-021-01703-5

Arunachalam, S., & Kothari, A. (2011). An experimental study of hindi and English perfective interpretation. *Journal of South Asian Linguistics*, *4*(1), 27–42. https://ojs.ub.unikonstanz.de/jsal/index.php/jsal/article/download/35/21/0.

Bánki, A., de Eccher, M., Falschlehner, L., Hoehl, S., & Markova, G. (2022). Comparing online webcam-and laboratory-based eye-tracking for the assessment of infants' audio-visual synchrony perception. *Frontiers in Psychology*, *6162*. DOI: https://doi.org/10.3389/fpsyg.2021.733933

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*. https://arxiv.org/abs/1406.5823. DOI: https://doi.org/10.18637/jss.v067.i01

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PloS one*, *3*(8). e3022. DOI: https://doi.org/10.1371/journal.pone.0003022

Bott, O., & Hamm, F. (2014). Cross-linguistic variation in the processing of aspect. In *Psycholinguistic approaches to meaning and understanding across languages* (pp. 83–109). Springer. DOI: https://doi.org/10.1007/978-3-319-05675-3_4

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. DOI: https://doi.org/10.7717/peerj.9414

Calabrich, S. L., Oppenheim, G. M., & Jones, M. W. (2021). Episodic memory cues in acquisition of novel visual-phonological associations: a webcam-based eye-tracking study. In *Proceedings of the annual meeting of the cognitive science society*, *43*. https://escholarship.org/uc/item/76b3c54t.

Chabal, S., Hayakawa, S., & Marian, V. (2022). Language is activated by visual input regardless of memory demands or capacity. *Cognition*, *222*, 104994. DOI: https://doi.org/10.1016/j.cognition.2021.104994

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, *90*, 123–133. DOI: https://doi.org/10.1016/j.jebo.2013.03.003

Chen, M. C., Anderson, J. R., & Sohn, M. H. (2001). What can a mouse cursor tell us more? correlation of eye/mouse movements on web browsing. In *Chi'01 extended abstracts on human factors in computing systems* (pp. 281–282). DOI: https://doi.org/10.1145/634067.634234

Chouinard, B., Scott, K., & Cusack, R. (2019). Using automatic face analysis to score infant behaviour from video collected online. *Infant Behavior and Development*, *54*, 1–12. DOI: https://doi.org/10.1016/j.infbeh.2018.11.004

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive psychology*. https://psycnet.apa.org/doi/10.1016/0010-0285(74)90005-X. DOI: https://doi.org/10.1016/0010-0285(74)90005-X

Dalmaijer, E. (2014). Is the low-cost eyetribe eye tracker any good for research? Tech. rep. PeerJ PrePrints. https://peerj.com/preprints/585v1.pdf. DOI: https://doi.org/10.7287/peerj.preprints.585v1

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12. DOI: https://doi.org/10.3758/s13428-014-0458-y

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a web browser? comparing response times collected with javascript and psychophysics toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12. DOI: https://doi.org/10.3758/s13428-015-0567-2

De Swart, H. (1998). Aspect shift and coercion. *Natural language & linguistic theory*, *16*(2), 347–385. DOI: https://doi.org/10.1023/A:1005916004600

de Swart, H. (2012). Verbal aspect. In *The oxford handbook of tense and aspect*. DOI: https://doi.org/10.1093/oxfordhb/9780195381979.013.0026

Degen, J., Kursat, L., & Leigh, D. (2021). Seeing is believing: testing an explicit linking assumption for visual world eye-tracking in psycholinguistics. Unpublished, retrieved 13/05/2021. https://github.com/thegricean/eyetracking_replications/blob/master/writing/2021_cogsci/sunbrehenyreplication.pdf.

Dow, S., Kulkarni, A., Klemmer, S., & Hartmann, B. (2012). Shepherding the crowd yields better work. In *Proceedings of the acm 2012 conference on computer supported cooperative work* (pp. 1013–1022). DOI: https://doi.org/10.1145/2145204.2145355

Dowty, D. R. (1979). *Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*. Dordrecht: Reidel. DOI: https://doi.org/10.1007/978-94-009-9473-7

Ehinger, B. V., Groß, K., Ibs, I., & König, P. (2019). A new comprehensive eye-tracking test battery concurrently evaluating the pupil labs glasses and the eyelink 1000. *PeerJ*, *7*, e7086. DOI: https://doi.org/10.7717/peerj.7086

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* (pp. 1–20). DOI: https://doi.org/10.3758/s13428-021-01694-3

Federico, G., & Brandimonte, M. A. (2019). Tool and object affordances: an ecological eye-tracking study. *Brain and cognition*, *135*, 103582. DOI: https://doi.org/10.1016/j.bandc.2019.103582

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). Labvanced: a unified javascript framework for online studies. In *International conference on computational social science (cologne)*.

Gagné, N., & Franzen, L. (2021). How to run behavioural experiments online: best practice suggestions for cognitive psychology and neuroscience. DOI: https://doi.org/10.31234/osf.io/nt67j

Gvozdanović, J. (2012). Perfective and imperfective aspect. In *The oxford handbook of tense and aspect*. DOI: https://doi.org/10.1093/oxfordhb/9780195381979.013.0027

Hauger, D., Paramythis, A., & Weibelzahl, S. (2011). Using browser interaction data to determine page reading behavior. In *International conference on user modeling, adaptation, and personalization* (147–158). Springer. DOI: https://doi.org/10.1007/978-3-642-22362-4_13

Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. C. (2017). Noiserobust fixation detection in eye movement data: Identification by two-means clustering (i2mc). *Behavior research methods*, *49*(5), 1802–1823. DOI: https://doi.org/10.3758/s13428-016-0822-1

Huang, J., White, R., & Buscher, G. (2012). User see, user point: gaze and cursor alignment in web search. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1341–1350). DOI: https://doi.org/10.1145/2207676.2208591

Huang, Y., & Snedeker, J. (2020). Evidence from the visual world paradigm raises questions about unaccusativity and growth curve analyses. *Cognition*, *200*, 104251. DOI: https://doi.org/10.1016/j.cognition.2020.104251

Huettig, F., & Altmann, G. T. M. (2011). Looking at anything that is green when hearing "frog": How object surface colour and stored object colour knowledge influence language-mediated

overt attention. *The Quarterly Journal of Experimental Psychology*, *64*(1), 122–145. DOI: https://doi.org/10.1080/17470218.2010.481474

Huettig, F., Guerra, E., & Helo, A. (2020). Towards understanding the task dependency of embodied language processing: the influence of colour during language-vision interactions. *Journal of Cognition*, *3*(1). DOI: https://doi.org/10.5334/joc.135

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. DOI: https://doi.org/10.1016/j.actpsy.2010.11.003

Jeschull, L. (2007). The pragmatics of telicity and what children make of it. In *Proceedings of the 2nd conference on generative approaches to language acquisition north america* (pp. 180–187). Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.492.5503&rep=rep1&type=pdf.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 ieee 12th international conference on computer vision* (pp. 2106–2113). IEEE. DOI: https://doi.org/10.1109/ICCV.2009.5459462

Jun, E., Hsieh, G., & Reinecke, K. (2017). Types of motivation affect study selection, attention, and dropouts in online experiments. *Proceedings of the ACM on Human-Computer Interaction 1(CSCW)* (pp. 1–15). DOI: https://doi.org/10.1145/3134691

Kochari, A. R. (2019). Conducting web-based experiments for numerical cognition research. *Journal of cognition*, *2*(1). DOI: https://doi.org/10.5334/joc.85

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, *13*(10), 1292–1298. DOI: https://doi.org/10.1038/nn.2635

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. DOI: https://doi.org/10.1371/journal.pone.0130834

Li, Q., Joo, S. J., Yeatman, J. D., & Reinecke, K. (2020). Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest. *Scientific reports*, *10*(1), 1–11. DOI: https://doi.org/10.1038/s41598-019-57204-1

Madden, C. J., & Zwaan, R. A. (2003). How does verb aspect constrain event representations? *Memory and Cognition*, *31*, 663–672. DOI: https://doi.org/10.3758/BF03196106

Madsen, J., Julio, S. U., Gucik, P. J., Steinberg, R., & Parra, L. C. (2021). Synchronized eye movements predict test scores in online video education. *Proceedings of the National Academy of Sciences*, *118*(5). DOI: https://doi.org/10.1073/pnas.2016980118

Magliano, J. P., & Schleich, M. C. (2000). Verb aspect and situation models. *Discourse processes*, *29*(2), 83–112. DOI: https://doi.org/10.1207/S15326950dp2902_1

Magnuson, J. S. (2019). Fixations in the visual world paradigm: where, when, why? *Journal of Cultural Cognitive Science*, *3*(2), 113–139. DOI: https://doi.org/10.1007/s41809-019-00035-3

Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition*, *108*(3), 866–873. DOI: https://doi.org/10.1016/j.cognition.2008.06.005

Martin, F., & Demirdache, H. (2020). Partitive accomplishments across languages. *Linguistics*, *58*(5), 1195–1232. DOI: https://doi.org/10.1515/ling-2020-0201

Martin, F., Demirdache, H., del Real, I. G.,Van Hout, A., & Kazanina, N. (2020). Children's non-adultlike interpretations of telic predicates across languages. *Linguistics*, *58*(5), 1447–1500. DOI: https://doi.org/10.1515/ling-2020-0182

Minor, S., Mitrofanova, N., & Ramchand, G. (2022a). Fine-grained time course of verb aspect processing. https://drive.google.com/file/d/1BUExbQzd2fbllbr80O9reJRmjaFZeFjB/view?usp=sharing. DOI: https://doi.org/10.1371/journal.pone.0264132

Minor, S., Mitrofanova, N., Guajardo, G., Vos, M., & Ramchand, G. (2022b). Temporal information and event bounding across languages: Evidence from visual world eyetracking. Talk at Semantics and Linguistic Theory 32. https://osf.io/tv3b8/.

Ooms, K., Dupont, L., Lapon, L., & Popelka, S. (2015). Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental setups. *JOURNAL OF EYE MOVEMENT RESEARCH*, *8*(1), 20. DOI: https://doi.org/10.16910/jemr.8.1.5

Ouzts, A. D., & Duchowski, A. T. (2012). Comparison of eye movement metrics recorded at different sampling rates. In *Proceedings of the symposium on eye tracking research and applications* (pp. 321–324). DOI: https://doi.org/10.1145/2168556.2168626

Papoutsaki, A., Laskey, J., & Huang, J. (2017). Searchgazer: Webcam eye tracking for remote studies of web search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval* (pp. 17–26). DOI: https://doi.org/10.1145/3020165.3020170

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). Webgazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the 25th international joint conference on artificial intelligence (ijcai)* (pp. 3839–3845). AAAI. https://par.nsf.gov/servlets/purl/10024076.

Peelle, J. E., & Van Engen, K. J. (2021). Time stand still: Effects of temporal window selection on eye tracking analysis. *Collabra: Psychology*, *7*(1), 25961. DOI: https://doi.org/10.1525/collabra.25961

R Core Team. (2019). R: A language and environment for statistical computing (version 3.6.1) [computer software]. r foundation for statistical computing. *Vienna, Austria*.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual review of vision science*, *2*, 437–457. DOI: https://doi.org/10.1146/annurev-vision-082114-035733

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 71–78). DOI: https://doi.org/10.1145/355017.355028

Saslow, M. G. (1967). Effects of components of displacement-step stimuli upon latency for saccadic eye movement. *Josa*, *57*(8), 1024–1029. DOI: https://doi.org/10.1364/JOSA.57.001024

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain sciences*, *10*(4), 251. DOI: https://doi.org/10.3390/brainsci10040251

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. DOI: https://doi.org/10.3758/s13428-017-0913-7

Slim, M., & Hartsuiker, R. (2021a). Online visual world eye-tracking using webcams. In *Architectures and mechanisms for language processing*. Paris, France. https://amlap2021.github.io/program/148.pdf.

Slim, M. S., & Hartsuiker, R. (2021b). Visual world eyetracking using webgazer.js. DOI: https://doi.org/10.31234/osf.io/5adgf

Slote, J., & Strand, J. F.(2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods*, *48*(2), 553–566. DOI: https://doi.org/10.3758/s13428-015-0599-7

Sogo, H. (2013). Gazeparser: an open-source and multiplatform library for low-cost eye tracking and analysis. *Behavior research methods*, *45*(3), 684–695. DOI: https://doi.org/10.3758/s13428-012-0286-x

Tafaj, E., Kasneci, G., Rosenstiel, W., & Bogdan, M. (2012). Bayesian online clustering of eye movement data. In *Proceedings of the symposium on eye tracking research and applications* (285–288). DOI: https://doi.org/10.1145/2168556.2168617

Tanenhaus, M. K., & Soivey-Knowlton, M. J., & Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. DOI: https://doi.org/10.1126/science.7777863

Van Hout, A. (2011). Past tense interpretations in dutch. In *Organizing grammar* (pp. 241–251). De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110892994

van Hout, A. (2018). On the acquisition of event culmination. *Semantics in language acquisition* (pp. 96–121). DOI: https://doi.org/10.1075/tilar.24.05hou

Vendler, Z. (1967). *Linguistics in philosophy*. Ithaca, N.Y.: Cornell University Press. DOI: https://doi.org/10.7591/9781501743726

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755* https://arxiv.org/abs/1504.06755.

Yang, W., Chan, A., Chang, F., & Kidd, E. (2020). Four-year-old mandarin-speaking children's online comprehension of relative clauses. *Cognition*, *196*, 104103. DOI: https://doi.org/10.1016/j.cognition.2019.104103

Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision Making*, *16*(6), 1486. http://journal.sjdm.org/21/210525/jdm210525.pdf.

Zhou, P., Crain, S., & Zhan, L. (2014). Grammatical aspect and event recognition in children's online sentence comprehension. *Cognition*, *133*(1), 262–276. DOI: https://doi.org/10.1016/j.cognition.2014.06.018