Behavior Research Methods (2023) 55:3786–3804
https://doi.org/10.3758/s13428-022-01989-z

Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js

Mieke Sarah Slim[1] · Robert J. Hartsuiker[1]

Accepted: 23 September 2022 / Published online: 2 November 2022

Abstract
The visual world paradigm is one of the most influential paradigms to study real-time language processing. The present study tested whether visual world studies can be moved online, using PCIbex software (Zehr & Schwarz, 2018) and the WebGazer.js algorithm (Papoutsaki et al., 2016) to collect eye-movement data. Experiment 1 was a fixation task in which the participants looked at a fixation cross in multiple positions on the computer screen. Experiment 2 was a web-based replication of a visual world experiment by Dijkgraaf et al. (2017). Firstly, both experiments revealed that the spatial accuracy of the data allowed us to distinguish looks across the four quadrants of the computer screen. This suggest that the spatial resolution of WebGazer.js is fine-grained enough for most visual world experiments (which typically involve a two-by-two quadrant-based set-up of the visual display). Secondly, both experiments revealed a delay of roughly 300 ms in the time course of the eye movements, possibly caused by the internal processing speed of the browser or WebGazer.js. This delay can be problematic in studying questions that require a fine-grained temporal resolution and requires further investigation.

Keywords Web-based eye tracking · Visual world paradigm · Online experiments · Language comprehension · Psycholinguistics

Introduction

Over the last decades, the visual world paradigm has proven to be one of the most fruitful techniques for studying real-time language processing (see Huettig et al., 2011, for review). The typical setup of a visual world experiment is relatively simple: Participants listen to auditory linguistic stimuli while they look at a display that contains visual stimuli (although the paradigm has also been used to test language production, e.g., Griffin & Bock, 2000). An eye-tracking device is used to track the eye movements of the participants. Since there is a tight temporal link between visual attention and language processing, this setup provides informative data.

This link was first observed by Cooper (1974), who let participants listen to short narratives while they looked at a display that contained nine pictures. Cooper's results showed that the participants tended to look at objects in the visual world that are related to the linguistic input that they are processing at that moment of time. For instance, the participants looked more often at a picture of a zebra upon hearing the word *zebra* in the narrative compared to when the word *zebra* is not mentioned. This effect emerged rapidly: People tended to fixate on the related object within 200 ms after the word onset (see also Matin et al., 1993; Saslow, 1967). Since Cooper's seminal findings, the visual world paradigm has been used to test real-time language processing at a wide range of linguistic levels, such as phonemic or phonological processing (e.g., Allopenna et al., 1998; Huettig & McQueen, 2007; Snedeker & Trueswell, 2004), syntactic processing (e.g., Altmann & Kamide, 1999; Kamide et al., 2003; Tanenhaus et al., 1995) or semantic and pragmatic processing (e.g., Degen & Tanenhaus, 2016; Huang & Snedeker, 2009, 2018; Sun & Breheny, 2020).

Even though the visual world paradigm has become one of the most fruitful and versatile paradigms for studying real-time language processing, it has an important limitation: It requires expensive and stationary eye-tracking equipment,

✉ Mieke Sarah Slim
mieke.slim@ugent.be

[1] Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

as well as a researcher that is present in the lab to administer the experiment. This makes eye-tracking-based experiments rather cumbersome, especially in comparison to other behavioral experiments that can be carried out remotely over the Internet (e.g., Gosling & Mason, 2015; Gibson et al., 2011; Pinet et al., 2017). The advantages of web-based testing are eminent: Because participants do not have to come to the lab, it is easy to recruit many participants (see Hartshorne et al., 2018, for a particularly successful attempt of recruiting a large sample size) or target a specific group of participants (e.g., speakers of a language that is not commonly spoken in the country in which the researcher resides). Moreover, since online experimentation does not require a researcher to supervise the experiment, data collection is also much faster and more efficient compared to lab-based experimentation.

Fortunately, there is hope that the visual world paradigm, as well as other eye-tracking-based paradigms, could be moved to the Internet. Most remote high-end eye-tracking devices use a near-infrared illumination to cause a reflection on the participants' cornea and a high-end video recorder to capture images of the eye. Image processing techniques then locate the pupil based on the visible reflection on the cornea, and use this information to estimate the participants' eye movements (e.g., SR Research, 2021; Tobii Pro, 2021). Recent software developments allow us to estimate gaze locations on videos with a lower resolution than those recorded with a high-end eye-tracking device (e.g., Valenti et al., 2009; Valliappan et al., 2020; Xu et al., 2015). In particular, Papoutsaki et al. (2016) developed a JavaScript-based library named *WebGazer.js*. WebGazer.js's algorithm consists of two main components: A pupil detector that looks for the position of the pupils in the webcam stream and a gaze estimator. This gaze estimator uses regression analysis to approximate the location of the looks on the screen. These regression analyses used by the gaze estimator are guided by the interactions of the participant, such as mouse clicks and cursor movements. WebGazer.js can therefore be used to collect eye-movement data in web-based experiments, but we know relatively little about the spatio-temporal resolution of the eye-movement data that WebGazer.js provides, or whether the data is accurate enough to use in psycholinguistic research.

The first studies that used WebGazer.js to conduct eye-tracking experiments showed promising results. In a lab-based experiment, Papoutsaki et al. (2018) used a high-end eye-tracking device (a Tobii Pro X2-120) and WebGazer.js in combination with a consumer-grade webcam to track participants' eye movements while they were typing. Their results showed that distinct eye-movement patterns can be distinguished for touch and non-touch typists. Moreover, these eye-movement patterns were found in both the data collected with the Tobii eye tracker and in the data collected with WebGazer.js. WebGazer.js was thus able to replicate data collected with a high-end eye tracker, although visualizations of this data do suggest that the quality of the WebGazer data is somewhat poorer relative to the Tobii data, showing more variance in both the spatial and the temporal domains.

In another study, Semmelmann and Weigelt (2018) tested the viability of web-based eye-tracking experiments using WebGazer.js and consumer-grade webcams by conducting an experiment that consisted of three tasks: (i) A fixation task in which the participants fixated on a circle that appeared on the screen for 2000 ms, (ii) a pursuit task in which the participants followed a circle that moved on the screen, and (iii) a free-viewing task in which the participants looked at a photograph of a face. Their results showed that WebGazer was suitable for all three tasks, although the spatial and temporal resolution was poorer compared to the standards of a high-end eye-tracking device. Looking at the spatial resolution of the data from the fixation and the pursuit tasks, the in-lab acquired data revealed an offset between the estimated fixation position and the stimulus of roughly 15–19% of the screen size. Nevertheless, Semmelmann and Weigelt's free-viewing task showed that WebGazer.js was able to replicate previous findings from lab-based experiments: The participants tended to fixate on the eyes when they look at an image of a face compared to other regions of interest (such as the mouth and the nose), which corroborates findings that Westerners tend to focus their attention at the eyes when they see a face (e.g., Blais et al., 2008). Regarding the temporal resolution of the data, the fixation task data showed that the saccade towards the stimulus started roughly 250–375 ms after stimulus onset and lasted 450–750 ms on average. Finally, it must be noted that Semmelmann and Weigelt's experiment showed considerable variance between participants. This variance is not only due to individual differences between participants, but also in terms of the hardware they used to do the experiment (e.g., quality of the webcam, stability of the Internet connection, lighting used in the room).

These few previous studies revealed that, in principle, web-based eye tracking can detect eye-movement patterns. However, these studies also reveal that the quality of the data is considerably weaker compared to data from lab-based experiments that used high-end eye trackers. Especially the temporal resolution observed in Semmelmann and Weigelt (2018) may be worrisome: Based on lab-based experimentations, we know that it takes roughly 200 ms to execute a saccade (e.g., Matin et al., 1993), whereas Semmelmann and Weigelt's results showed that it took roughly 750 ms until participants settled their gazes on a stimulus. This raises the question of whether webcam-based techniques are suitable for conducting visual world experiments (or behavioral research in general), which often requires a precise temporal resolution.

In the present study, we gain insight into this question by conducting two web-based experiments that used WebGazer.js to track participants' eye movements. Experiment 1 was a fixation task that was inspired by Semmelmann and Weigelt's (2018) fixation task. In this experiment, the participants looked at a fixation cross that appeared on the screen for 1500 ms. The aim of this experiment was to gain further insight into the spatial and temporal resolution of web-based eye-tracking data, and to see how this data quality is influenced by properties of the hardware, such as webcam sample rate or the calibration threshold of the webcam eye tracker (more below).

Experiment 2 was a web-based replication of a visual world study that was previously conducted by Dijkgraaf et al. (2017) in an in-lab setting using a high-end eye-tracker. This experiment investigated predictive processing based on verb meaning in sentence comprehension. Predictive processing based on verb information is a finding that is often observed in visual world studies: If the visual display contains only one picture of an object (e.g., a picture of a letter) that would be a fitting argument following the verb in the auditory stimulus (e.g., a sentence like *Mary read a letter*), participants already tend to fixate on the fitting picture during the processing of the verb, before the onset of the second noun (see Altmann & Kamide, 1999; Kamide et al., 2003; Borovsky et al., 2012; Hintz et al., 2017, 2020, *inter alia*). By replicating Dijkgraaf et al.'s visual world experiment in a web-based setting, and by comparing our web-based data with their lab-based data, we will get a rich insight into the viability of the use of WebGazer.js in visual world studies.

Experiment 1: Fixation task

Participants

This experiment was approved by the Ethics Committee of the Faculty of Psychological and Educational Sciences at Ghent University. All participants gave informed consent by selecting a check box on one of the first web pages in the experiment, before the task started.

We tested 57 native speakers of English via Prolific (https://www.prolific.co/), who were paid £1.25 for their participation. Although the task was non-linguistic, and therefore did not necessarily require native speakers of English, we set these screening restrictions so that the participants were comparable to those of Experiment 2 (which did require native speakers of English). Prior to the experiment, the participants were instructed to not wear glasses during the experiment, and none of the participants reported to have worn glasses in a post-experimental questionnaire. Finally, all participants opened the experiment in the Google Chrome Desktop browser. They were instructed to open the experiment on Google Chrome, and then were not able to continue to the experiment unless they indicated that they were on Google Chrome in a pre-experimental question.

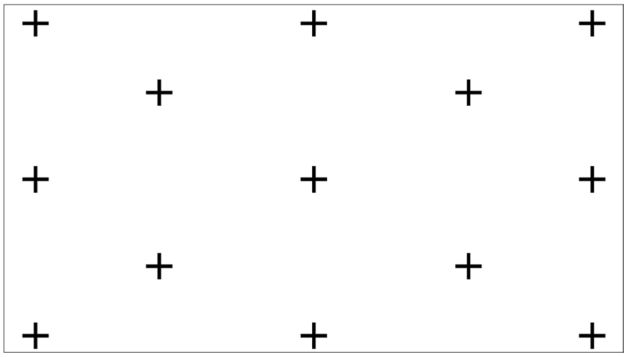

Fig. 1 The 13 positions in which the stimuli could appear on the screen in Experiment 1

The results of two participants were not saved on the server due to connectivity issues, and therefore we could not include these participants the data analysis. Thus, the data of 55 participants were included in the final analyses.

Stimuli materials

The participants looked at a fixation cross that appeared in one of thirteen positions on the screen (Fig. 1). An important difference between remote web-based eye tracking and in-lab eye tracking is that the computer screens of participants in web-based studies vary in sizes and resolution, whereas the participants of a lab-based eye-tracking experiment usually all carry out the experiment on the same hardware. To ensure that the experiment appeared at least similar for all participants, we set the size and position of the fixation cross relative to the screen size of the participant: the height (and width) of the cross is 15% of the height of the participant's screen (for example, if a participant's screen had a resolution of 1440 by 640, the size of the fixation cross is 96 by 96 pixels, because the experiment was shown in full screen). Note that four of the 13 positions were in the middle of each quadrant of the screen, which are typically the positions where the images are shown on the visual display in a visual world experiment. The other nine positions were spread out over the screen. The addition of these positions allowed us to do a more fine-grained calculation of estimated gaze-location accuracy, which may be beneficial for future studies. The fixation cross appeared six times at each of the 13 positions, resulting in 78 trials. Trial order was fully randomized for each participant.

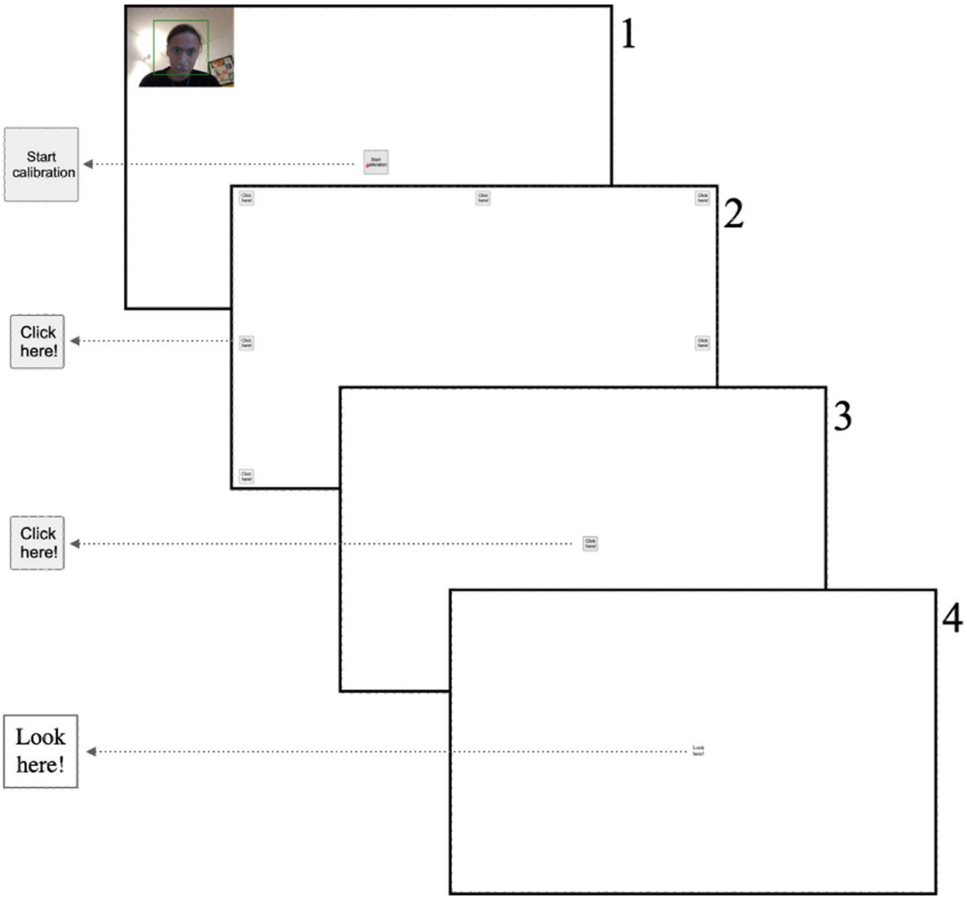

Fig. 2 The four steps of the calibration procedure. In the first step, the participants need to position themselves in front of their webcams by positioning the image of their head in the green square. They can then start the calibration by pushing the button in the middle of the screen (which says 'Start calibration'). In the second step, they click on eight buttons (that all say 'Click here') that are placed along the edges of the screen, while they follow the cursor closely with their eyes. After they have clicked on all eight buttons, a new button appears in the center of the screen (step 3). After they have clicked this button, step 4 will begin. In this final step, the participants fixate on the middle of the screen (marked by the words 'Look here!') for 3 s. During this step, the calibration score is calculated

Procedure

The task was carried out remotely over the Internet and implemented using PennController for Ibex (*PCIbex*; Zehr & Schwarz, 2018). PCIbex has an eye-tracker element that uses WebGazer.js to estimate the locations of the participants' eye gazes. The code that was used to implement this experiment is available at GitHub (https://github.com/MiekeSlim/Moving-visual-world-experiments-online)[1].

The experiment started with a welcome page, an informed consent form, and instructions on how to set up the webcam. On the welcome page, the participants were asked to give the browser permission to use the webcam. Here, they were informed that we are not recording any videos, but only save information about where they look on the screen. Moreover, we instructed the participants to ensure that they were correctly seated in front of the webcam and that they were in a well-lit environment. Once the participant went through the welcome screens, the browser was prompted to switch to full screen.

The participants then continued to a calibration procedure, which consisted of four steps (Fig. 2). First, participants saw the webcam stream along with a green frame indicating the required head position, so they could position themselves correctly in front of their webcam (step 1). Then, the participants clicked on eight buttons that were placed along the edges of their screen (step 2). Once they had clicked on all these buttons, a new button appeared in the center of the screen (step 3). The participants clicked on

[1] The code and resources on GitHub can be easily imported into the PCIbex Farm (https://farm.pcibex/net), which can be used to host PCIbex experiments. Instruction on how to import these files on the PCIbex farm are provided in the GitHub repository (and in the PCIbex documentation: https://doc.pcibex.net/how-to-guides/github/).

this button and then fixated on the center of the screen (indicated by the phrase ‘Look here!’) for 3 s (step 4). During these 3 s, the webcam eye tracker calibrated by calculating the proportion of estimated looks on the center of the screen. This proportion of estimated looks (ranging from 0 to 100%) indicated the success of calibration (the *calibration score*).

The required calibration threshold was set at 5% (meaning that the calibration is considered successful if the eye tracker calculated that at least 5% of the estimated looks fell on the center of the screen in the last step of the calibration procedure). If this threshold was not met, the participant went through the full calibration procedure again. Because one of the goals of Experiment 1 was to test the influence of calibration score on the accuracy of the eye-tracking data, we wanted to obtain a wide range of calibration scores. Therefore, we chose to use a low calibration threshold in Experiment 1.

The fixation task started after successful calibration. Each trial started with a 500-ms black fixation cross in the center of the screen. The target fixation cross then appeared in one of the 13 positions of the screen for 1500 ms and then the next trial started automatically. In case the participant exited the full screen modus of the browser (e.g., by hitting the escape key), the browser was prompted to switch to full screen at the end of each trial. Every 13 trials were interspersed with a calibration trial. The participants clicked on a button that appeared in the middle of the screen, and then fixated on the middle of the screen for three seconds so that the eye tracker can calibrate (steps 3 and 4 of the calibration procedure). If more than 5% of the estimated looks fell outside the target region in the center of the screen, the participant went through the full calibration again. If calibration was successful, the next trial started automatically.

Analyses and results

Data treatment

The data and analysis scripts are online available at: https://osf.io/yfxmw/. Experiment 1 collected two types of gaze measurements: screen coordinates of the estimated gaze location and the quadrant on which each gaze was directed. The screen coordinates are given as pixel coordinates. This is not a uniform measure because the location of pixels on the screen depends on the participant’s screen size and resolution. Therefore, we first standardized these coordinates by defining the position of the estimated gaze as a percentage of the participants’ screen width and height (e.g., the coordinate of the pixel in the center of the screen is defined as (50, 50), regardless of the participants’ screen resolution).

We then aggregated the data into 100-ms bins, so that we had the same number of observations per participant and per trial. Here, however, we noticed that the duration of each eye-tracker recording did not always have the expected length of 1500 ms (mean duration 1456 ms; sd 122 ms; range 801–2210 ms). Longer recordings are likely due to short lags in the experiment. For instance, browser glitches (possibly caused by poor browser performance, i.e., the speed by which the browser renders and executes the functions prompted by the experiment script) may lengthen the trials/recordings by several milliseconds. Shorter recordings, on the other hand, seem to be due to the sample frequency of the participant’s webcam (see below): In some cases, the webcam only recorded one frame per several 100 ms. In these cases, the eye tracker did record for the full 1500 ms, but the last recorded frame came in well before 1500 ms. It is noteworthy that both the particularly long and short recordings were observed in the same (few) participants, which suggests that both the longer and the shorter duration of the eye-tracking recording may be caused by the same underlying problem, most likely browser processing speed and/or webcam quality. We did not remove any trials or participants prior to our analyses due to this issue since the variation between the participants and their hardware are relevant for these analyses. However, we did remove the data of each bin that was above 1500 ms (484 of 4279 recorded bins, i.e., 11.31% of the total number of recorded bins) to create more homogeneity in our dataset.

Finally, for each time bin, we calculated the Euclidean distance between the estimated gaze location and the center of the stimulus, using the following formula:

$$\sqrt{\left(X_{stimulus\ location} - X_{estimated\ gaze\ location}\right)^2 + \left(Y_{stimulus\ location} - Y_{estimated\ gaze\ location}\right)^2} \quad (1)$$

The analyses of the Euclidean distance between the stimulus location and the estimated gaze location will give information about both the temporal resolution of the data (since we expect this distance to become smaller over time) and the spatial resolution of the data (since this Euclidean distance expresses the offset of the estimated gaze location and the actual stimulus location).

Analyses of Euclidean distance

We first tested whether the Euclidean distance between the gaze location and the stimulus position changed over the time course of the trials. Due to the explorative nature of this experiment, we mostly relied on visual inspection of the data in this part of the analyses. Looking at the visualization of the time course of the data (Fig. 3A), the start of each trial is characterized by a saccade towards the stimulus location, which can be identified as a decrease in the Euclidean distance between the stimulus position and the estimated gaze location. On average, this decrease in distance started at roughly 200 ms after the stimulus onset. After 500 ms, the distance between the estimated fixation

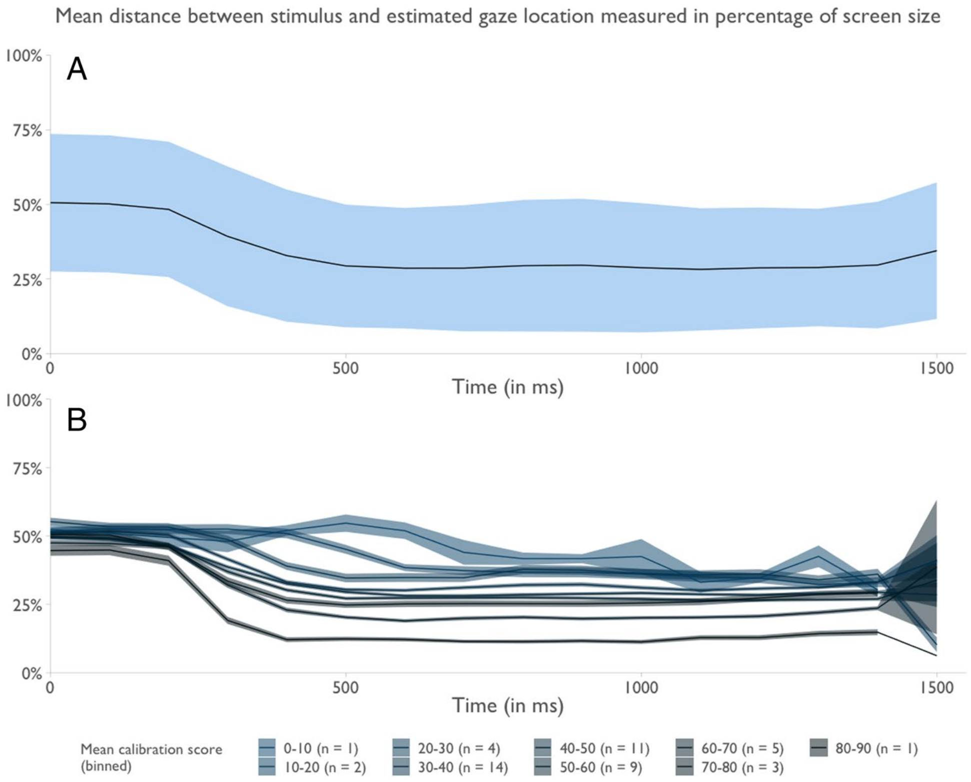

Fig. 3 The mean Euclidean offset, measured in percentage of screen size, over the duration of the trials. **A** Mean Euclidean offset averaged across all participants. Here, the *blue ribbon* represents the standard deviation, to represent the large distribution of the data. **B** Mean Euclidean offset divided into separate mean calibration bins. In **B**, the ribbons represent the standard error, since showing the standard deviations would have made the figure difficult to interpret. Note that the standard errors in **B** increase substantially towards the end of the trial. Here, the number of observations decrease because the recordings of some trials are shorter than 1500 ms, which results in larger standard errors (see Section 1.6.1.)

location and the stimulus position remains relatively stable over time, which suggests that the participant's fixations were settled on the stimulus. Within this fixation time window of 500–1500 ms, the mean distance between the estimated gaze location and the center of the stimulus was still roughly 30% of the screen size.

The data show considerable variability, as can be seen in the large standard deviation visualized in Fig. 3A. Possibly, this variability can partly be explained in terms of calibration success. We plotted the mean gaze location over time again, but now split up the data following the mean calibration scores of the participants in ten-point bins (Fig. 3B; resulting in a total of nine bins, since no participant had a higher mean calibration score of 90). This plot suggests that calibration score affects both the spatial and the temporal accuracy of the data. First, we see that the estimated gaze locations are closer to the stimulus for participants with a higher mean calibration score compared to those with a lower mean calibration score. Second, it also seems that the fixation time window (in which the estimated gaze locations are settled on the stimulus) starts earlier for the participants who scored higher calibration scores in general. Note that the number of participants in each bin is unbalanced (see the legend of Fig. 3B).

In addition, we investigated whether the position of the stimulus on the screen influences the spatial accuracy. Here, we analyzed the spatial resolution of the data in the time window between 500 and 1500 ms after the stimulus onset, because the participants fixations settled on the stimulus roughly 500 ms after the stimulus onset (Fig. 3). Visual inspection of the data (Fig. 4) suggests that the webcam eye tracker can discriminate between the quadrants of the screen. However, it seems that the accuracy is better if the stimulus is displayed in the center of the screen and less accurate if the stimuli are presented in the far corners of the screen (as was the case in the bottom-left, top-left, bottom-right, and top-right positions). It is worth noting that previous studies have also shown that professional-grade eye-tracking devices also estimate the gaze location more precisely if the participant looks at the center, rather than on the edge, of the screen, e.g., Ehinger et al., 2019).

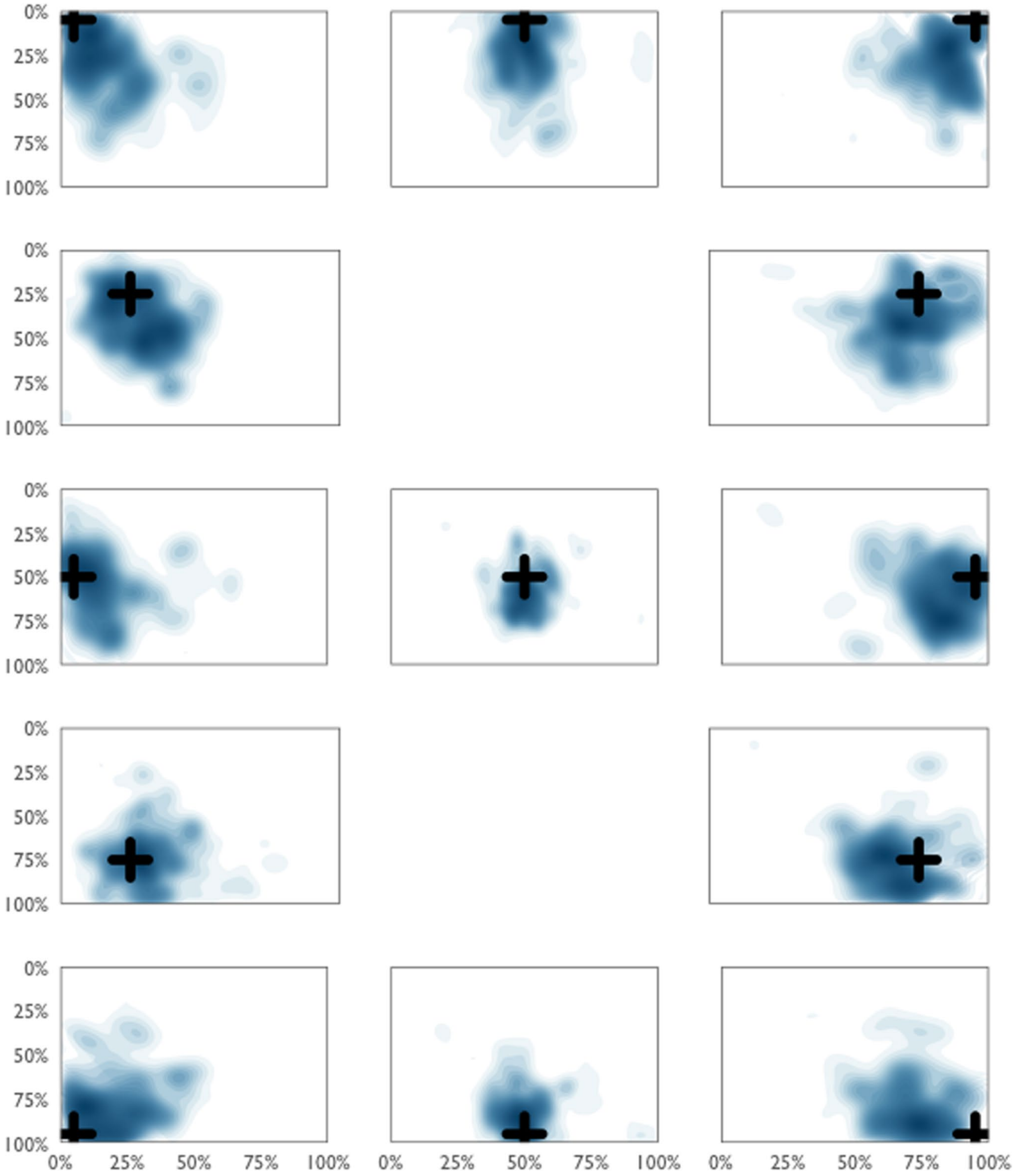

Fig. 4 The density of looks over the screen in the fixation time window (500–1500 ms) broken down in all 13 fixation cross positions. Note that each panel represents the full screen. The stimulus positions that correspond to the center of each quadrant are shown in the second and fourth row. The *black crosses* show the center of the fixation target positions

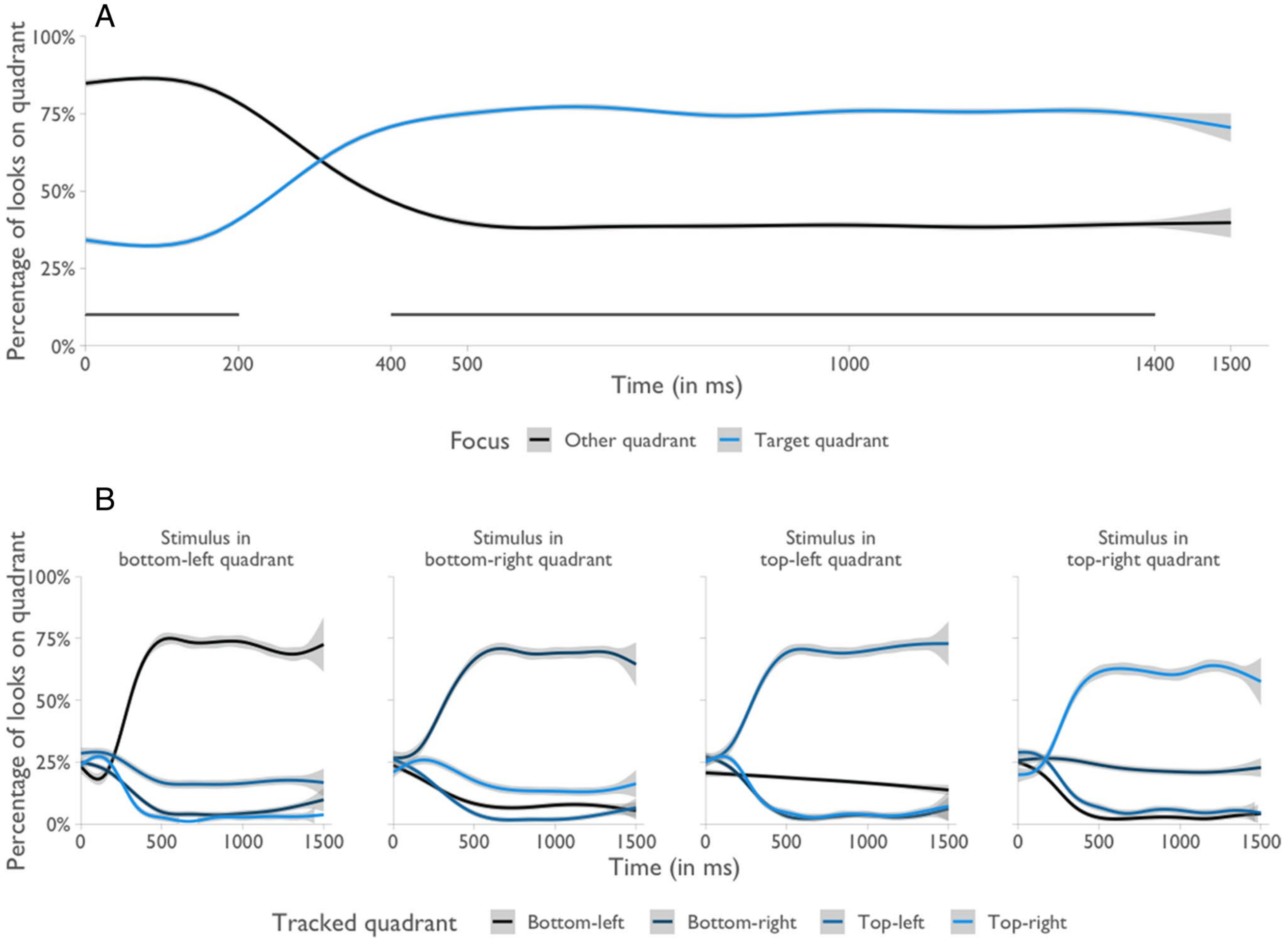

Fig. 5 The proportion of looks on the target quadrant increases after 200 ms, and becomes significantly higher than looks at any of the other three quadrants after 400 ms (**A**). In addition, WebGazer.js is able to discriminate looks on any of the four quadrants, and the highest proportion of non-target looks are directed to the quadrant that is either above or below the target quadrant (**B**). Note that this figure only shows data of the four-quadrant based positions

Analyses of the quadrant-based data

We also recorded the screen quadrant in which each estimated gaze location fell. Here, our analyses focused on the central locations of each quadrant (0.25, 0.25–0.25, 0.75–0.75, 0.25–0.75, 0.75). This type of data is relevant in the context of a visual world experiment, which typically involves the presentation of four images in the quadrants of the screen.

We started these analyses by comparing looks on the target quadrant to looks on the other three quadrants (Fig. 5). This way of data coding allowed us to perform explorative inferential statistics over the time course of the trial. We binarized the data (per trial and participant, a bin was coded as 1 if more than 30% of looks fell on the target quadrant, otherwise it was coded as 0). These data were analyzed with a cluster permutation analysis to identify temporally adjacent time bins that showed a significant difference in the likelihood of looks on the target quadrant and the likelihood of looks on the other quadrants ($p < 0.05$). Each bin was tested for significance using a logit mixed-effect model (which contained random intercepts for Subject and Position), and then, the data were randomly permuted and tested for significance again. This latter step was repeated 10000 times to create an empirical null distribution. Finally, the empirical distributions were compared to the differences in the observed clusters, to test the reliability of the differences in the observed clusters (e.g., Huang & Snedeker, 2020; Maris & Oostenveld, 2007). This cluster permutation analysis was conducted using the *permutes* package in *R* (Voeten, 2021).

This analysis showed two clusters of adjacent time bins in which there was a reliable difference in the likelihood of target quadrant fixations compared to fixations on the other quadrants (Fig. 5A). First, between 0 and 200 ms, participants were more likely to look at the other three quadrants

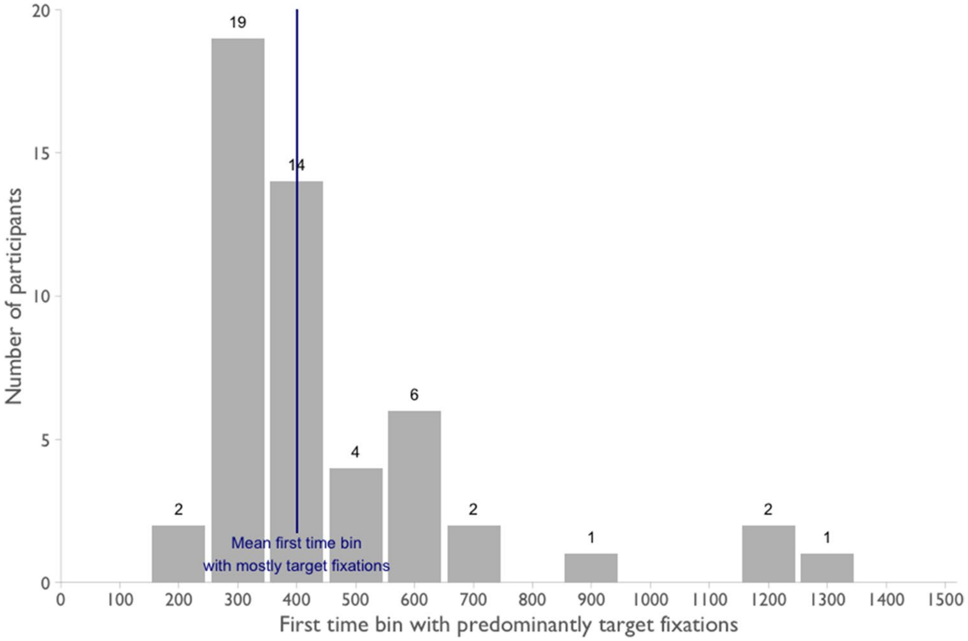

Fig. 6 The distribution of the first time bin with a higher mean proportion of target fixations than non-target fixations is skewed: There is larger variation across the participants who, on average, settle their gazes on the target later than 400 ms after stimulus onset (which is the mean) than before 400 ms after stimulus onset

than on the target quadrant. Second, between 400 and 1400 ms, they were more likely to look at the target quadrant than on any other quadrant. This suggests that, on average, participants settled their gazes 400 ms after the stimulus onset, although participants may launch their fixations as early as 200 ms after stimulus onset.

In addition, it is worth noting that there were no substantial differences in the spatio-temporal accuracy across the four quadrants (Fig. 5B), although the proportion of fixations is descriptively slightly lower if the stimulus was presented in the top-right quadrant compared to any of the three other quadrants. Moreover, most of the estimated looks that were recorded as non-target looks after 400 ms fell on the quadrant above or below the target quadrant. This suggests that WebGazer.js is better in discriminating left-right looks from top-down looks (Fig. 5B).

Finally, we tested the variation across participants in the temporal domain by calculating, for each participant, the first time bin in which there was a higher proportion of estimated looks on the target quadrant than on any of the other quadrants on average. This time bin was not observed for four participants, because their estimated looks fell mostly on the non-target quadrants on average throughout the time course of the trials. For the other 51 participants, this time bin fell between 200 and 1300 ms of the stimulus onset (Fig. 6).

Figure 6 shows that the distribution of the first time bins in which the estimated gaze locations predominantly fell on the target quadrant across participants is skewed: This time bin was observed before 400 ms for 21 participants, at 400 ms for 14 participants, and between 500 ms and 1300 ms for the other 16 participants (and again, not even observed for four other participants).

The influence of calibration

Altogether, our data shows considerable variation across participants. In Section 2.4.2, we briefly hinted that this variation can partly be explained in terms of calibration score: Descriptively, the estimated gaze locations of participants who obtained a higher calibration score seemed to be more temporally and spatially accurate (Fig. 3B). Here, we briefly report some additional calculations on the influence of eye tracker calibration on the quality of the data. A more elaborate description of these calculations can be found in Appendix A. First, we observed that calibration score correlated with the participants' webcam quality (expressed in the number of frames that are recorded each second, the so-called *fps* or *frames per second* rate; $\rho = 0.852$, $p < 0.001$). This indicates that the eye tracker can calculate a more precise calibration score if there are more recorded frames. Second, we also observed that calibration score correlated with the spatial accuracy of the data. Focusing on all 13 stimulus positions, we observed that the Euclidean offset between the estimated gaze locations and the stimulus was larger for participants who obtained lower calibration scores on average in the fixation time window (500–1500 ms; $\rho = -0.472$, $p < 0.001$). Taking the data from the four center-quadrant positions, we observe a similar correlation between mean calibration score and proportion of target quadrant looks in the fixation time window ($\rho = 0.395$, $p = 0.002$). This suggests that calibration score also influences the spatial accuracy of the data if proportion of quadrant looks are measured, rather than more fine-grained estimate gaze coordinates.

Finally, we also observed a correlation between calibration score and the temporal accuracy of the data. Focusing on the data of the four quadrant-based positions alone, we observed that the first time bin with predominantly target fixations was later on average for participants with lower calibration scores than for participants with higher calibration scores ($\rho = 0.747$, $p < 0.001$). Recall that there were four participants for who this first time bin of target fixation was not observed. These participants were not included in this correlation test.

Altogether, the analyses of Experiment 1 showed that it takes roughly 400–500 ms until WebGazer.js detects that the participants' gaze settled on the stimulus location. In addition, we observed that both the temporal and the spatial accuracy seemed to be better for participants who obtained higher calibration scores. Therefore, we raised the calibration threshold to 50 in Experiment 2, which is reported below.

Experiment 2: Replication of Dijkgraaf et al. (2017)

Participants

This experiment was approved by the Ethics Committee of the Faculty of Psychological and Educational Sciences at Ghent University. All participants gave informed consent by selecting a check box on one of the first web pages in the experiment before the task started.

Based on the finding from Experiment 1 that data quality improves with a higher calibration score, we set the calibration threshold at 50. However, some participants were not able to (consistently) reach this threshold in the calibration procedure (more below). These participants were redirected to another experiment that did not involve webcam eye-tracking (not reported here).

We pre-set our desired sample size at ninety participants, following a rule-of-thumb to recruit three times the size of the original sample size of Dijkgraaf et al.'s (2017) experiment ($n = 30$). We recruited 330 native speakers of English via Prolific. We redirected 240 participants to another experiment, because they did not reach the calibration threshold in five attempts. The remaining 90 participants all took part in the visual world experiment. We did not exclude any of these participants in the analyses. All participants were paid £4.50 for their participation.

Stimulus materials

The materials and design were identical to the English monolingual version of the experiment reported by Dijkgraaf et al. (2017). The experiment involved 18 experimental trials and 18 filler trials. In all these trials, the participants listened to a recording of a sentence while looking at a display of four pictures (arranged over the four quadrants of the screen). The experimental trials were presented in two conditions: the *constraining* and the *neutral* condition. In the constraining condition, only one of the four pictures depicted an appropriate post-verbal object (e.g., a letter following the verb *read*). In the neutral condition, on the other hand, all four pictures displayed appropriate post-verbal objects (e.g., a letter, a backpack, a car, and a wheelchair following the verb *steal*; Fig. 7).

In the filler trials, the display contained either no, two, or three pictures that depicted appropriate post-verbal arguments. All pictures are black-and-white line drawings that were taken from a normed database constructed by Severens, Van Lommel, Ratinckx, and Hartsuiker (2005). Dijkgraaf et al. (2017) matched the object names of the pictures for frequency, phoneme count, and syllable count across the conditions.

The sentences were simple four-word active transitive sentence (e.g., *Mary reads a letter*). The subject phrase was the same in all trials (*Mary*). The object noun always started with an indefinite article. The object noun always started with a consonant, so the article could not serve as a prediction cue. The sentences were pronounced by a female native speaker of Dutch, who speaks English as a second language (and majored in English linguistics and literature at university). Dijkgraaf et al.'s (2017) original study did not only involve monolingual English participants, but also Dutch–English bilingual participants. They selected this speaker for the sentence recordings because of her clear pronunciation of both English and Dutch sentences. Her accent was rated by native speakers of English as 5.3 on a seven-point scale (where 1 = "very foreign accented" and 7 = "native speaker"). We decided to re-use these recordings, to keep our replication as close as possible to Dijkgraaf et al.'s original experiment.

The 18 experimental stimuli and 18 filler trials were divided into two stimulus lists (named *A* and *B*). We took the lists from Dijkgraaf et al. (2017), who in turn assigned the experimental and filler trials pseudorandomly to the two lists, with the constraint that two sentences that belonged to the same stimulus set were not put in the same list. Each list contained nine constraining trials, nine neutral trials, and nine filler trials. Each trial contained a unique verb, but the

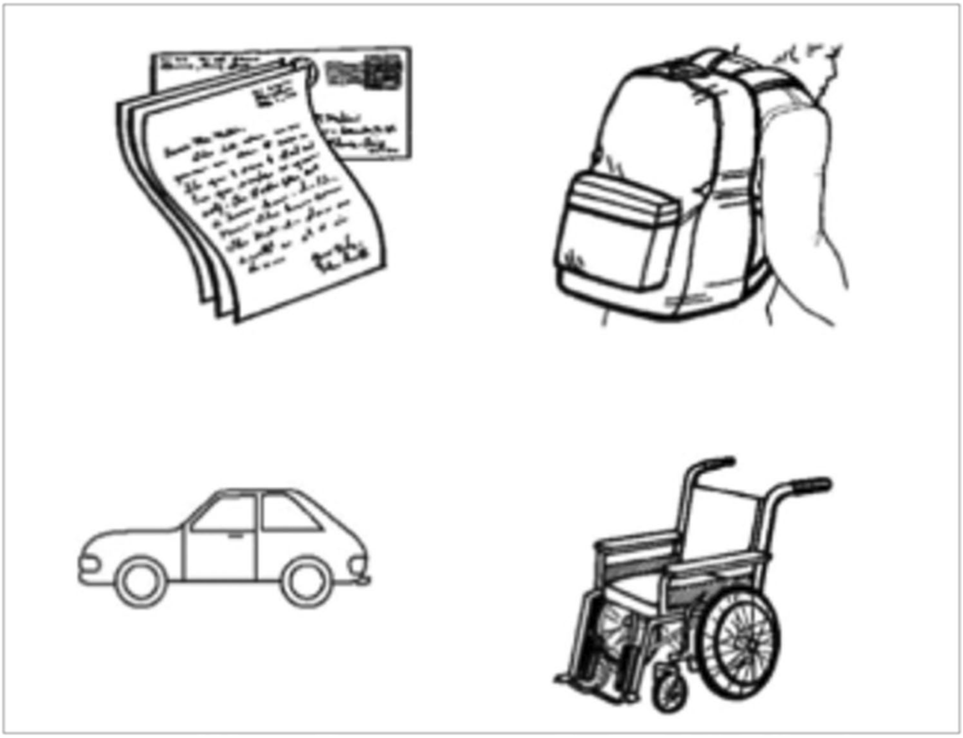

Fig. 7 Example of a visual scene used in Dijkgraaf et al. (2017). In the *constrained condition*, the sentence that accompanied this display was *Mary read a letter*. In the *neutral condition*, the sentence *Mary steals a letter* was played

displays were repeated across blocks. Within each list, the trials were fully randomized for each participant. The order of the lists was counterbalanced across participants.

Procedure

Like Experiment 1, the experiment was implemented using the PennController for Ibex library (Zehr & Schwarz, 2018). The code of this experiment is freely available on GitHub (https://github.com/MiekeSlim/Moving-visual-world-experiments-online)[2]. Also like Experiment 1, Experiment 2 started with general information and an informed consent pages. Following these pages, the browser switched to full screen and the webcam calibration procedure started. The calibration procedure was similar to the one in Experiment 1, with the difference that the calibration threshold was now set to 50. If this threshold was not met, the last step of the calibration procedure was repeated (in which the participant clicked on a button in the middle of the screen and then fixated at the center of the screen for 3 s). If the threshold of 50 was not met in five attempts, the participant was redirected to another online experiment that did not involve webcam eye tracking (not reported here).

After successful calibration, the participants listened to a sentence recording, so they could adjust the volume of their computer. The participants could replay this recording as often as needed. Once they indicated that they had set their volume, another sentence recording was played once. The participants typed in this sentence, so that we could check whether the participant had indeed set up their audio properly (which was the case, because all participants correctly typed in the sentence). These two sentence recordings were used as practice trials in Dijkgraaf et al.'s (2017) original experiment, but not in the present experiment.

Following the audio setup, a brief practice block of two trials was presented. Afterwards, the participants started the first block of the experiment by clicking on a 'start' button. After 18 trials (i.e., the first list of stimuli), the participants could take a short break. The second block of the experiment started with the presentation of an audio recording again, so the participants could check whether their volume was still set correctly. Then, all trials in the second list were presented in the second half of the experiment.

Each trial started with a calibration check, which consisted of step four of the calibration procedure (Fig. 2). If the threshold of 50 was not met, the full calibration

[2] The code and resources on GitHub can be easily imported into the PCIbex Farm (https://farm.pcibex/net), which can be used to host the experiment. Instructions on how to import these files on the PCIbex farm are provided in the GitHub repository.

procedure was repeated. If the threshold was met, the trial started automatically. As in Dijkgraaf et al. (2017), the four-picture display was shown for 2200 ms before the sentence recording started playing. Unlike Dijkgraaf et al., however, the display remained on the screen for an additional 500 ms after the auditory stimulus stopped playing. We added this short overspill time window because Experiment 1 suggested that WebGazer.js tends to be slower in recording the participants' eye movements than a remote high-end eye tracking device (see also Semmelmann & Weigelt, 2018). After this 500-ms overspill time, the next trial started automatically. Similar to Experiment 1, if a participant exited the full screen modus, the browser was prompted to full screen after each trial.

Data treatment and analyses

The data and analysis scripts are online available at https://osf.io/yfxmw/. In Experiment 2, we did not record the screen coordinates of the estimated gaze locations. Instead, we only recorded on which quadrant of the screen each estimated gaze location fell (like Dijkgraaf et al., 2017). Dijkgraaf et al. aggregated their data in 50-ms time bins. However, if we were to aggregate the online-acquired data in 50-ms time bins, we would create some empty bins due to the low sampling frequency of some participants' webcams. Instead, we aggregated the data in 100-ms bins, and applied this to Dijkgraaf et al.'s original data as well.

Dijkgraaf et al. (2017) tested the time course of the effect of Condition (*neutral* vs. *constrained*) on the proportion of target fixations by first determining a critical time window for analyses through visual inspection of the data. Then, each time bin was separately analyzed by modeling the likelihood of fixation on the target quadrant using generalized mixed-effect models (e.g., Jaeger, 2008). In these models, the dependent variable was the proportion of target fixation (transformed using the empirical logit formula, Barr et al., 2013). Condition (*neutral* vs. *constrained*) was included as a fixed effect, and Subject and Item were included as random effects (both slopes and intercepts).

We, however, used cluster permutation analysis (e.g., Hahn et al., 2015; Huang & Snedeker, 2020; Maris & Oostenveld, 2007). This analysis procedure, which we described in Experiment 1, has some important advantages over the procedure used by Dijkgraaf et al. (2017): It is less sensitive to the multiple-comparisons problem, and it is not needed to set an a priori time window, since cluster permutation analysis investigates adjacent clusters of statically reliable effects and then tests whether these clusters are statistically sound or whether they have occurred by chance. We conducted these cluster permutation analyses on both Dijkgraaf et al.'s data and on our online-acquired data. The analyses were done in *R* (R Core Team, 2021), using the *permutes* package (Voeten, 2021).

Results and discussion

The effect of constraining verbs: In-lab vs web-based data collection

Dijkgraaf et al. (2017) analyzed the time window between 350 ms after verb onset and 200 ms after noun onset. Their time course analyses revealed that the effect of Condition first showed significance in the 450–500-ms after verb onset time bin.

Our cluster permutation test revealed that the difference between the *neutral* and the *constrained* condition was significant in the time window between 500 and 1300 ms after verb onset ($p < 0.001$; Fig. 8, top panel). In this time window, the proportion of fixations on the target image was higher in the *constrained* condition than in the *neutral* condition. Crucially, this time window starts before the mean onset of the second noun, which indicates that these looks are predictive looks based on the action denoted in the verb, as also observed by Dijkgraaf et al. (2017). Note that our re-analysis revealed that the significant time window started 50 ms later than in Dijkgraaf et al.'s original analyses. This is most likely a consequence of our choice to aggregate the data in 100-ms bins rather than in 50-ms bins.

The cluster analysis on our online-acquired data also showed an effect of Condition, although this effect emerged later than in the in-lab acquired data from Dijkgraaf et a. (2017): The cluster analysis revealed a significant effect of Condition in the time window between 700 and 1600 ms after the verb onset (Fig. 8, bottom panel; $p < 0.05$). This indicates that the overall effect observed in Dijkgraaf et al. (2017) is replicated, but the significant time window starts 200 ms later in the online-acquired data than in the lab-acquired data.

The online-acquired data in Experiment 2 thus seem to show a time lag compared to the lab-acquired data from Dijkgraaf et al. (2017). Therefore, we descriptively analyzed variation across participants in the time course of the recorded eye movements by calculating the first bin where, on average, there was a higher proportion of looks at the target quadrant than on any of the other three quadrants (aggregated across both conditions; similar to the analysis reported in Experiment 1; Fig. 9). This analysis showed considerable variation across participants: On average, the 1500–1600-ms time bin was the first time bin in which most estimated looks fell on the target picture than on any of the other pictures. However, this time bin ranged between 300 and 2000 across participants. This analysis thus suggests that the time lag observed in the online-acquired data of Experiment 2 is (at least partly) caused by individual variation across participants. We will return to this point in the General Discussion.

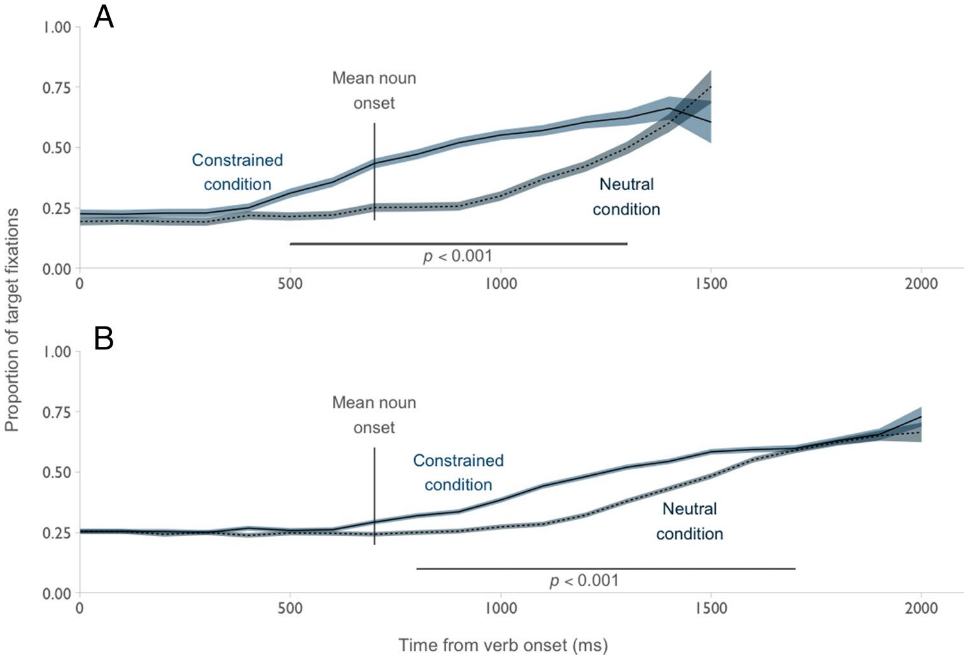

Fig. 8 Results from Dijkgraaf et al.'s (2017) lab-based study (**A**) and our web-based replication (**B**). Our web-based study replicated the global pattern observed by Dijkgraaf et al.: Participants looked at the target image earlier in the constrained condition than in the neutral condition. However, there is a delay in the time course of the online-acquired data compared to Dijkgraaf et al.'s data: In Dijkgraaf et al.'s original study, this effect emerged 500 ms after the verb onset, whereas this effect emerged 700 ms after the verb onset in the online-acquired data

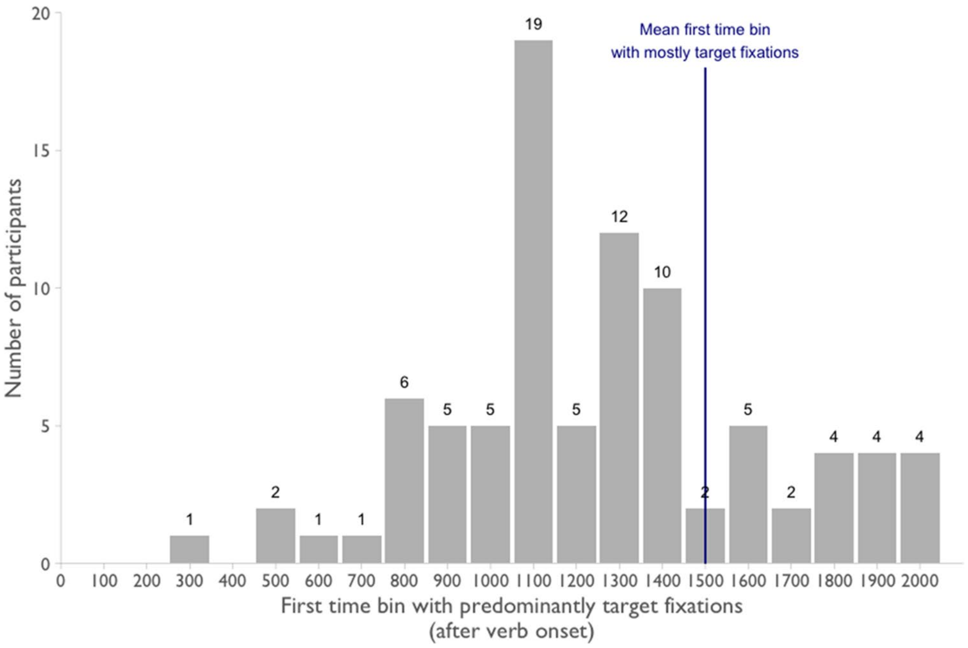

Fig. 9 The distribution of onset of the first time bin per participants in which there is, on average, a higher proportion of target fixations than non-target fixations seems to be somewhat skewed

The role of calibration

Here, we briefly report calculations that tested whether variation in the (spatio-temporal) accuracy of the data can be explained by calibration scores. More detailed descriptions of these calculations are reported in Appendix A.

First, we tested whether calibration correlated with the spatial accuracy of the data, by taking the subset of data recorded 1600 ms after verb onset until the end of the trial (i.e., the time window in which the webcam eye-tracker detected that the participants settled their gaze on the target picture, because there was no effect of Condition anymore). A Spearman's rank correlation test showed a weak and non-significant negative correlation between the proportion of target fixations and mean calibration scores (ρ = – 0.197, $p = 0.063$). Second, we tested variation across participants in the time course of the estimated gaze locations by calculating the first bin where, on average, there was a higher proportion of estimated looks at the target quadrant than on any of the other three quadrants (aggregated across both conditions). A Spearman's rank correlation test revealed no correlation between the mean obtained calibration score and the onset of the average first time bin of target fixation (ρ = 0.116, $p = 0.282$).

These analyses suggest that variation in calibration score in Experiment 2 did not seem to correlate with the spatial or temporal resolution of the eye-tracking data. This contrasts with Experiment 1, where the calibration threshold was much lower (5 instead of 50; a point we will return to in the General discussion).

Effect size and power considerations

Descriptively, the difference between the proportion of looks on the target in the *Constrained* and *Neutral* condition seems to be smaller in the online-acquired data than in the in-lab-acquired data from Dijkgraaf et al. (2017), suggesting that the effect size is smaller in the online-acquired data. This observation has important implications for the required sample size for web-based eye tracking compared to lab-based data, since the required sample size to obtain adequate statistical power relies on the size of the effect (e.g., Brysbaert & Stevens, 2018; Green & MacLeod, 2016). We tested the difference in effect size in both the online-acquired and lab-acquired data by taking the data from the time window that showed a significant effect of Stimulus Condition (as revealed by our cluster permutation analysis: 700–1700 ms in the online-acquired data and 500–1300 ms in the lab-acquired data). We then modeled the likelihood of target fixations with a logit mixed-effect model (which contained Stimulus Condition as a (sum-coded) fixed effect, and random intercepts for Participant and Sentence), which was constructed using the *lme4* package in R (Bates et al., 2015). As expected, this model showed that participants were more likely to fixate on the target quadrant during this time window in the *constrained* than in the *neutral* condition in both the online-acquired data (*beta* = 0.164, $z = 3.983$, $p < 0.001$) and in the lab-acquired data (*beta* = 0. 274, $z = 3.50$, $p < 0.001$). The beta coefficients of these models provide an estimate of the effect size, and this coefficient is considerable smaller in the online-acquired data than in the lab-acquired data (i.e., 0.164 vs. 0.274, which is roughly 60% of the effect observed in-lab).

To test whether Experiment 2 was indeed sufficiently powered given this finding, we conducted an explorative simulation-based power calculation on Dijkgraaf et al.'s (2017) data using the *mixedpower* package in R (Kumle et al., 2021), in which we tested the number of participants required to obtain an effect that is half the size of the effect observed in Dijkstra et al.'s original data. This analysis was a simulation-based power calculation in which we simulated the data in the significant time window 1000 times for different numbers of participants (the full details of this power analyses are given in Appendix B). Each dataset was tested for significance using the same logit mixed-effect model procedure given above. This power analysis showed that we reach sufficient power (i.e., 80% or higher) to detect an effect half the size observed in Dijkgraaf et al.'s data with 70 to 75 participants. This suggests that our experiment may be slightly overpowered (as also indicated by the small standard errors in Fig. 8; it is also worth noting that the observed effect size in our replication was slightly bigger than half the size of in Dijkgraaf et al.'s observed effect).

Altogether, it seems that (i) the effect size in a web-based visual world study is roughly half as observed in an in-lab experiment (which should be considered in determining the required sample size for a web-based eye-tracking study), and (ii) collecting two to two-and-a-half times as many participants as an in-lab study eye-tracking study may be sufficient to obtain sufficient power. However, since the power analysis conducted here is explorative and the size of the effect may depend on calibration threshold, design of the display, or population, these observations require further validation.

General discussion

This study aimed to gain insight in the viability of web-based visual-world eye-tracking experimentation using the WebGazer.js algorithm in combination with consumer-grade webcams to track participants' eye movements. In Experiment 1, we tested the spatial and temporal resolution of the webcam eye tracker in a simple fixation task. The experiment revealed that it took roughly 400 to 500 ms until the participants settled their gaze on the stimulus. Once they fixated on the stimulus, the mean offset between the stimulus and the estimated gaze location was still roughly 30% of the screen size. The spatial and temporal accuracy, however, improved with calibration score. In Experiment 2, we replicated a visual world study from Dijkgraaf et al. (2017) that tested predictive processing based on verb information in language comprehension. Experiment 2 replicated the overall pattern observed in Dijkgraaf et al.'s original data (which was acquired in a lab-based setting): Participants tended to fixate earlier on a picture displaying the object argument if the verb was semantically constraining. However, we observed a delay in the latencies of the eye movements: The anticipatory looks to the object image surfaced on average roughly 200 ms later in our online-acquired data compared to Dijkgraaf et al.'s lab-acquired data. Below, we will discuss the implications of these findings for the efficacy of web-based visual world experiments.

Spatio-temporal accuracy of online-acquired eye movement data

In Experiment 2, we conducted a web-based replication of a visual world experiment from Dijkgraaf et al. (2017). This experiment looked at effects of predictive processing based on verb information, which is an often-observed effect in visual world studies (e.g., Altmann & Kamide, 1999) and also observed in Dijkgraaf et al.'s data. The results of our web-based replication in Experiment 2 mirrored the overall pattern observed in Dijkgraaf et al.: Participants looked at the image depicting the post-verbal argument earlier in case the verb was semantically constraining compared to if the verb was not constraining towards one of the four pictures on the display. However, the onset of this effect emerged roughly 200 ms later in our online-acquired data compared to Dijkgraaf et al.'s in-lab acquired data.

Firstly, this finding reveals that the spatial resolution of the webcam eye-tracker is accurate enough to discriminate fixations across the four quadrants of the screen, which is needed for most visual world experiments. In Experiment 1, we observed that the spatial accuracy improved with increasing calibration score, but Experiment 2 (with a calibration threshold of 50) found no such relation. However, the data of Experiment 2 still showed variation across participants, which resulted in a smaller effect size in our web-based replication (Experiment 2) compared to Dijkgraaf et al.'s (2017) original lab-based study. This suggests that web-based eye-tracking is suitable for visual displays with four images that are arranged in quadrants (or less), but a larger sample size is required for web-based than for in-lab experimentation. Moreover, our results indicate that web-based eye tracking is less suitable for visual displays and paradigms that require a fine-grained spatial resolution, like visual search paradigms, eye-tracking-while-reading, or visual world paradigms that test small effects on more crowded displays.

Secondly, the data of both Experiment 1 and Experiment 2 revealed a latency in the expected time to execute a saccade. In Experiment 1, this latency could partly be explained in terms of calibration success: Participants who obtained lower calibration scores showed a delay in fixation onset. In Experiment 2, however, we did not observe a clear relation between calibration and temporal resolution of the eye-tracking data. It is worth noting that the delay observed in WebGazer.js data seems to be systematic. It was also observed by Semmelmann and Weigelt (2018), who used WebGazer.js to track eye movements in a fixation task similar to the one in Experiment 1. They observed that it took roughly 600 ms to execute a saccade, which is 400 ms longer than typically observed (e.g., Matin et al., 1993). Moreover, a direct comparison between the performance of a high-end eye-tracking device and WebGazer.js conducted by Papoutsaki et al. (2018) also descriptively suggested a delay in the WebGazer.js data compared to the data from the high-end eye-tracker.

The time lag in the temporal resolution seems to be an artefact of the web-based nature of our experiment, especially because there is no plausible reason to assume that the cognitive mechanisms involved in language processing and visual attention systematically differ depending on whether participants are tested in the lab or from home. In fact, numerous studies have shown that web-based testing is a viable alternative to lab-based testing to study language processing (e.g., Gibson et al., 2011; Hartshorne et al., 2018; Hilbig, 2016; *inter alia*). Therefore, we think that the delay may be caused by two separate factors. First, the internal processing speed of the WebGazer.js algorithm and/or the rendering speed of the browser may be slower compared to the software that high-end eye-trackers use to estimate the gaze locations. Second, individual variation across participants and the contexts in which they participate (e.g., hardware, attendance, environment, etc.) may produce outliers, which causes a skewed distribution. This skewed distribution can cause a delay of the effect in the overall data patterns. In Experiment 1, we observed that this variation correlated with calibration scores; Experiment 2 used a calibration threshold which reduced the impact of variation in calibration. However, other factors such as attendance or certain environmental factors could still have influenced the accuracy of Experiment 2. Future work is required to characterize these factors and test how they can be filtered out, which could reduce the delay in the online-acquired eye-tracking data.

This delay and noise in the temporal resolution seems to be the biggest challenge for web-based visual world experiments because this paradigm is often used to study questions about the fine-grained time course of real-time language processing. The problematic impact of this systematic delay becomes clear in the results of Experiment 2.

This experiment studied *anticipatory effects* in eye movements, which are characterized as the tendency to look at a target image prior to the onset of the targeted linguistic fragment (in our case, the noun in the object phrase of the auditory stimulus). However, due to the delay in the observed time course, the assumed predictive looks were not detectable until after the onset of the object noun. (However, we assume that they are, given the global similarity in pattern in Dijkgraaf et al.'s (2017) original data and our online-acquired data). Therefore, the results of Experiment 2 thus show that care is required in interpreting the time course of online-acquired eye-tracking data. Therefore, web-based eye-tracking in its current form may not be suitable to study questions that require a fine-grained temporal resolution, such as effects in a small time window.

Variation across participants

The previous subsection already discussed that part of the inaccuracy in online-acquired eye-tracking data seemed to be caused by the variation across participants, which can partly be attributed to differences in the hardware used by the participants. In both Experiments 1 and 2, we found that the accuracy of the eye-tracker (measured in terms of calibration scores) was influenced by the sampling frequency of the webcam.

As hinted in the previous subsection, there are many more aspects that can influence eye-tracker calibration and data accuracy, such as ambient lighting, facial characteristics, environmental distractions, or other individual differences between participants. This variation can be somewhat reduced through clear and explicit instructions: Participants should not only be instructed to move their head as little as possible and that they are centrally seated in front of their screen and webcam, but they also need to make sure that they are in a well-lit room with few distractions, and that they should look at the screen throughout the experiment. However, there is reason to think that the performance of the webcam eye-tracker is also dependent on unalterable facial characteristics. Papoutsaki et al. (2018), for instance, observed that the face detector used by WebGazer.js recognizes faces with a lighter skin tone more often than faces with a darker skin tone, which is a bias that is more often observed in facial recognition systems (e.g., Coe & Atay, 2021). Encouragingly, there seems to be a trend in software development to reduce biases in automated facial recognition software (e.g., Atay et al., 2021; Lunter, 2020), so these biases will hopefully be reduced in the future.

Moreover, note that the influence of calibration on the spatio-temporal accuracy was only detectable in Experiment 1, whereas it was less clear in Experiment 2. We already attributed this difference in findings to two factors: (i) In Experiment 1, we measured spatial accuracy in terms of screen coordinates of the estimated fixation instead of larger region of interests, and (ii) the calibration threshold in Experiment 2 was set at 50% whereas that in Experiment 1 was set at 5%. This suggests that our calibration threshold of 50% successfully served as a filter to reduce (part of) the variation across participants. However, we still observe considerable variation across the participants in Experiment 2, both in terms of spatial and temporal accuracy. As also mentioned in the previous subsection, this is not a surprise because many differences between participants are not filtered out by calibration score. A participant who looks away from the screen during the trial, for instance, may still go through calibration successfully. Do note that it is inherent to web-based testing that there is less control over the participants' attendance and environment compared to lab-based studies, but these issues may be minimized by clear instructions, attention checks, and frequent calibration checks (which requires the participant to look at the screen).

Recommendations and further studies

Experiments 1 and 2 revealed that web-based visual world eye-tracking using the participants' webcams may be a suitable alternative to lab-based testing, at least if the study does not require a very fine-grained spatio-temporal resolution. Here, we will give several recommendations for future inquiries on web-based eye-tracking.

As already mentioned, we do not recommend web-based eye-tracking in its current state for paradigms that require fine-grained spatial data, like visual search, reading-based paradigms, visual world paradigms with more than four items on the display, or visual world studies that test small effects. This is because the spatial accuracy of WebGazer.js is considerably poorer than that of high-end remote eye-trackers.

However, the biggest threat to the viability of web-based visual world studies using WebGazer.js may be the temporal resolution of the data. We observed a lag in the time course of online-acquired data, and we would therefore not recommend web-based eye tracking for temporally sensitive data (like visual world paradigms with a short time window of interest). However, we also believe that there are reasons to be optimistic for the future of web-based eye tracking. Once the nature of this time lag is more precisely defined, we could take it into account in the processing of online-acquired eye-tracking data. Moreover, this information could also be used to improve web-based eye-tracking techniques (see Yang and Krajbich, 2021). Despite our optimism about the future of web-based eye tracking, we emphasize that care is required in using current web-based eye-tracking methods to study questions that require a precise temporal resolution.

Additionally, we can give recommendations regarding calibration threshold based on the experiments reported here. As mentioned above, there was no correlation between calibration score and spatio-temporal accuracy in Experiment 2 in contrast to Experiment 1. This suggests that the calibration threshold in Experiment 2 (50) successfully reduced variation that influenced calibration across participants. A disadvantage of using a threshold is that many participants needed to be excluded: In Experiment 2, we needed to recruit 330 participants to obtain a sample size of 90 participants. This may be a discouraging finding at first glance. However, we should note that we did not do any efforts (besides giving the instructions again) to improve the calibration scores for participants who consistently failed calibration. Now, however, we know that webcam sampling rate was a major influence on calibration threshold. Therefore, the number of participants that do not meet the calibration threshold could be reduced by testing the webcam sampling rate before calibration. If webcam fps rate is poor, the participant could be instructed on how to improve the sampling rate (e.g., change devices if possible). Otherwise, if the participant is unable to improve the sampling rate, they can be excluded without having to go through the full calibration procedure. However, we should note that many other factors besides fps rates may affect the calibration scores and many participants may not be able to change devices. Therefore, we do recommend to clearly communicate to the participants that they may not be able to do the experiment, and/or have them redirected to another non-eye-tracking experiment.

In addition, the results of Experiment 2 suggested that we may not have needed to recruit as many participants as we did: We collected data of 90 participants, whereas explorative post hoc power analyses suggested that 75 participants would have been sufficient (Appendix B). We encourage authors of future web-based replications to test the difference in effect size in web-based visual world studies compared to that in in-lab studies because these data could help further improve recommendations for sample size in web-based visual world experiments.

Conclusions

Across two experiments, we tested the efficacy of web-based visual world studies using WebGazer.js in combination with the participants' own webcams. We firstly observed that the spatial accuracy of the web-based eye tracker is accurate enough to discriminate looks across the four quadrants of a computer screen. Secondly, we observed a delay in the latency of the eye-movement data compared to what we would expect (based on previous studies). We hypothesized that this delay is due to variation across participants and the browser performance and/or internal processing speed of WebGazer.js. The spatial resolution of the webcam eye tracker therefore seems to be accurate enough for many visual world studies, since the typical display of a visual world study contains four items arranged in quadrants over the screen. However, at its current state, web-based eye tracking is not ideal for studying questions that require a close temporal resolution, given the time lag in the eye-movement data, although future inquiries may reduce the observed time lag. Nevertheless, the ease and efficiency in collecting data online makes web-based eye tracking an ideal technique for studying questions that may not require a very fine-grained spatio-temporal resolution).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-022-01989-z.

Acknowledgements We would like to thank Aster Dijkgraaf for sharing the materials and data from Dijkgraat et al. (2017) and Jeremy Zehr for his extensive help in programming the experiments. This research was funded by a PhD fellowship grant from Research Foundation Flanders, allocated to Mieke Sarah Slim (Grant number 53755).

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.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264.

Atay, M., Gipson, H., Gwyn, T., & Roy, K. (2021). Evaluation of Gender Bias in Facial Recognition with Traditional Machine Learning Algorithms. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–7). IEEE.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software, 67*(1). 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. https://doi.org/10.1371/journal.pone.0003022

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*(4), 417–436.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1).

Coe, J., & Atay, M. (2021). Evaluating impact of race in facial recognition across machine learning and deep learning algorithms. *Computers, 10*(9), 113.

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*.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science, 40*(1), 172–201.

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

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.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass, 5*(8), 509–524.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877–902.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*, 274–279.

Hahn, N., Snedeker, J., & Rabagliati, H. (2015). Rapid linguistic ambiguity resolution in young children with autism spectrum disorder: Eye tracking evidence for the limits of weak central coherence. *Autism Research, 8*(6), 717–726.

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition, 177*, 263–277.

Hilbig, B. E. (2016). Reaction time effects in lab-versus web-based research: Experimental evidence. *Behavior Research Methods, 48*(4), 1718–1724.

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(9), 1352.

Hintz, F., Meyer, A. S., & Huettig, F. (2020). Visual context constrains language-mediated anticipatory eye movements. *Quarterly Journal of Experimental Psychology, 73*(3), 458–467.

Huang, Y., & Snedeker, J. (2020). Evidence from the visual world paradigm raises questions about unaccusativity and growth curve analyses. *Cognition, 200*, 104251.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology, 58*(3), 376–415.

Huang, Y. T., & Snedeker, J. (2018). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology, 102*, 105–126. https://doi.org/10.1016/j.cogpsych.2018.01.004

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57*(4), 460–482.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*(1), 133–156.

Kumle, L., Võ, M. L. H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543.

Lunter, J. (2020). Beating the bias in facial recognition technology. *Biometric Technology Today, 2020*(9), 5–7.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics, 53*(4), 372–380.

Papoutsaki, A., Gokaslan, A., Tompkin, J., He, Y., & Huang, J. (2018). The eye of the typer: A benchmark and analysis of gaze behavior during typing. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 16.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 3839–3845.

Pinet, S., Zielinski, C., Mathôt, S., Dufau, S., Alario, F.-X., & Longcamp, M. (2017). Measuring sequences of keystrokes with jsPsych: Reliability of response times and interkeystroke intervals. *Behavior Research Methods, 49*(3), 1163–1176.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for statistical computing. https://www.R-project.org/

Saslow, M. (1967). Effects of components of displacement-step stimuli upon latency for saccadic eye movement. *Josa, 57*(8), 1024–1029.

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465.

Severens, E., Van Lommel, S., Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta Psychologica, 119*(2), 159–187.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*(3), 238–299.

SR Research (2021). *EyeLink® 1000 Plus Brochure*. https://www.sr-research.com/wp-content/uploads/2018/01/EyeLink-1000-Plus-Brochure.pdf

Sun, C., & Breheny, R. (2020). Another look at the online processing of scalar inferences: An investigation of conflicting findings from visual-world eye-tracking studies. *Language, Cognition and Neuroscience, 35*(8), 949–979.

Tanenhaus, M. K., Spivey-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.

Tobii Pro. (2021). *Tobii* Pro Spectrum Product Description. https://www.tobiipro.cn/siteassets/tobii-pro/product-descriptions/tobii-pro-spectrum-product-description.pdf/?v=2.4

Valenti, R., Staiano, J., Sebe, N., & Gevers, T. (2009). Webcam-based visual gaze estimation. *International Conference on Image Analysis and Processing*, 662–671.

Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., et al. (2020). Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications, 11*(1), 1–12.

Voeten, C. C. (2021). Analyzing time series data using clusterperm. *Lmer*.

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.

Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision Making*, *16*(6), 1486.

Zehr, J., & Schwarz, F. (2018). *PennController for internet based experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832.

Open practices statement

Both Experiment 1 and 2 experiment were preregistered. These preregistrations can be accessed at https://osf.io/yfxmw/registrations. The data reported in this paper are available on https://osf.io/yfxmw/. Finally, a walkthrough of the experiment scripts (which can be imported into the freely available PCIbex Farm (https://farm.pcibex.net) is available at https://github.com/MiekeSlim/Moving-visual-world-experiments-online. The materials used in these studies are widely available.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.