Journal of Child Language (2024), 1–34
doi:10.1017/S0305000924000175

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Assessing two methods of webcam-based eye-tracking for child language research

Margaret Kandel and Jesse Snedeker

Department of Psychology, Harvard University, USA
Corresponding author: Margaret Kandel; Email: mkandel@g.harvard.edu.

(Received 10 June 2023; revised 20 February 2024; accepted 21 April 2024)

Abstract

We assess the feasibility of conducting web-based eye-tracking experiments with children using two methods of webcam-based eye-tracking: automatic gaze estimation with the WebGazer.js algorithm and hand annotation of gaze direction from recorded webcam videos. Experiment 1 directly compares the two methods in a visual-world language task with five to six year-old children. Experiment 2 more precisely investigates WebGazer.js' spatiotemporal resolution with four to twelve year-old children in a visual-fixation task. We find that it is possible to conduct web-based eye-tracking experiments with children in both supervised (Experiment 1) and unsupervised (Experiment 2) settings – however, the webcam eye-tracking methods differ in their sensitivity and accuracy. Webcam video annotation is well-suited to detecting fine-grained looking effects relevant to child language researchers. In contrast, WebGazer.js gaze estimates appear noisier and less temporally precise. We discuss the advantages and disadvantages of each method and provide recommendations for researchers conducting child eye-tracking studies online.

Keywords: web-based experimentation; eye-tracking; phonemic cohort competition; language comprehension; WebGazer

Introduction

Visual-world eye-tracking is an important tool for studying real-time language processing in children. In the visual-world paradigm, participants are presented with a display, and their eye-movements are recorded as they listen to or produce an utterance. Individuals systematically look to referents or associates of the words they hear (e.g., Cooper, 1974; Tanenhaus et al., 1995) or are planning to produce (e.g., Griffin & Bock, 2000; Meyer et al., 1998). Saccades are tightly linked to linguistic information, with fixations to relevant stimuli rising within 200ms of the onset of linguistic cues in adults (e.g., Allopenna et al., 1998; Cooper, 1974). This relationship has allowed researchers to use eye-movements to investigate a variety of questions in language processing (see Huettig et al., 2011 for review). This paradigm is particularly useful for child research, as it provides a non-invasive, real-time measure of language processing that doesn't require meta-linguistic reasoning (cf. grammaticality judgments, lexical decision), reading ability (cf. self-paced reading), or a lengthy set-up (cf. electroencephalography). Children similarly look to relevant stimuli shortly after the onset of linguistic cues, and visual-world experiments have been used with children to study multiple levels of language processing, including phonological (e.g., McMurray et al., 2018; Sekerina & Brooks, 2007), morphological (e.g., Özge et al., 2022; Zhou et al., 2014), syntactic (e.g., Contemori et al., 2018; Snedeker & Trueswell, 2004; Trueswell et al., 1999), semantic (e.g., Borovsky et al., 2012; Brouwer et al., 2019), and pragmatic processing (e.g., Cooper-Cunningham et al., 2020; Huang & Snedeker, 2009; Kampa & Papafragou, 2020).

Visual-world experiments are primarily conducted in university labs where researchers employ specialized equipment to monitor participant gaze (e.g., SR Research, 2021; Tobii, 2021). More recently, however, algorithms that determine gaze location based on webcam video have increased interest in conducting eye-tracking experiments without specialized equipment and outside of lab settings (e.g., Erel et al., 2022; Fraser et al., 2021; Papoutsaki et al., 2016; Valenti et al., 2009; Valliappan et al., 2020; Xu et al., 2015). Webcam-based eye-tracking allows researchers to conduct experiments over the internet, in either supervised settings (with an experimenter present over video conferencing) or unsupervised settings (with no experimenter present). Web-based testing has several advantages, many of which are particularly relevant to child research. Participants can complete experiments from the comfort of their own homes, where children may feel more at ease. This frees families from needing to travel to the lab and make babysitting arrangements for siblings. Unsupervised web-based experiments allow for even more efficient data collection, as sessions can occur outside of working hours at whatever time is most convenient for families. Collecting data over the internet gives researchers access to more diverse populations (see Henrich et al., 2010 for the importance of sample diversity) and languages not spoken near their home institutions. Webcam-based eye-tracking can also be used in conjunction with direct participant contact, allowing researchers to set up mobile labs wherever they can bring a laptop (e.g., schools, parks, museums, etc.).

Of the algorithms that track eye-gaze from webcam videos, the JavaScript library *WebGazer.js* (hereafter "WebGazer"; Papoutsaki et al., 2016) has garnered the most attention from behavioral researchers. WebGazer is open-source and has been integrated into popular frameworks for running online behavioral tasks, such as PCIbex (Zehr & Schwarz, 2018), JsPsych (de Leeuw, 2015), and Gorilla (Anwyl-Irvine et al., 2020). Gaze estimation occurs locally in the user's web-browser, and no video is saved, thus maintaining participant privacy. Although initially designed to detect eye-gaze during user interactions with webpages (Papoutsaki et al., 2016), recent studies have explored WebGazer's suitability for behavioral research with adults.

The results of these investigations are promising. WebGazer detects looks to perceptual stimuli shortly after they appear (e.g., Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2022) and has been used to replicate previously-observed eye-tracking effects in a variety of domains, including visual inspection of faces (Semmelmann & Weigelt, 2018), decision making (X. Yang & Krajbich, 2021), and language processing (Degen et al., 2021; Slim & Hartsuiker, 2022; Vos et al., 2022). However, WebGazer has limitations compared to the eye-tracking devices typically used for in-lab studies. Specifically, the offset between estimated gaze and stimulus locations is greater and looking patterns are delayed relative to in-lab studies (e.g., Degen et al., 2021; Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2022). At present, it is not clear to what extent this noise is attributable to WebGazer itself as opposed to properties of the less controlled web-based setting (e.g., variations in software, hardware, environments, and internet connections) or differences in participant behavior when completing studies online.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

Given these findings with adults, it seems reasonable to consider using WebGazer for web-based psycholinguistic studies with children. However, it is not obvious that WebGazer would perform as well when estimating child gaze. Child faces are smaller than those of adults, and children are likely to be in a different position relative to the webcam because of their height, which could reduce the accuracy of WebGazer's pupil detection and gaze estimation algorithms. In addition, young children are less likely to remain in the same position for the duration of a task, and they are unlikely to have the patience to sit through extensive calibration/recalibration procedures that improve accuracy in adult studies (e.g., Semmelmann & Weigelt, 2018; X. Yang & Krajbich, 2021). In fact, even high-end in-lab eye-trackers are less accurate when used with children (Dalrymple et al., 2018). Furthermore, children may have more difficulty maintaining attention when completing an experiment from home, where there may be more distractions than in controlled lab settings.

In the present study, we investigate whether it is possible to run web-based visual-world studies with school-aged children. We test two webcam eye-tracking methods: automatic gaze estimation with WebGazer and frame-by-frame annotation of gaze direction (e.g., Snedeker & Trueswell, 2004) from webcam videos recorded via Zoom teleconferencing software (https://zoom.us/). Experiment 1 directly compares these two methods in a visual-world language task with five to six year-old children. We assess how well these methods discriminate both robust fixation patterns (looks to target stimuli) as well as more subtle eye-movement patterns of the kind relevant to child language researchers (phonemic cohort competition effects; e.g., Allopenna et al., 1998; Sekerina & Brooks, 2007). By collecting both forms of gaze data simultaneously, we can assess the extent to which any noise observed in the WebGazer data stems from WebGazer itself as opposed to participant behavior or the web-based setting. Experiment 2 focuses more specifically on WebGazer, assessing its performance with child participants aged four to twelve years in a visual-fixation task. Experiment 2 was run without an experimenter present, allowing us to assess the feasibility of conducting unsupervised web-based eye-tracking studies with child participants.

Experiment 1: visual-world task

Experiment 1 comprised two linked experiments focused on the phonemic cohort competition effect. This effect is well-suited for testing the efficacy of web-based visual-world eye-tracking, as it has been replicated many times with both adults (e.g., Allopenna et al., 1998; Dahan & Gaskell, 2007; Dahan et al., 2001; Farris-Trimble & McMurray, 2013; Magnuson et al., 1999; inter alia) and children (e.g., Desroches et al., 2006; Sekerina & Brooks, 2007; Rigler et al., 2015; Weighall et al., 2017; inter alia), and the presence of cohort activation is often used to investigate higher-level linguistic constraints on incremental language processing (e.g., Dahan & Tanenhaus, 2004; Gaston et al., 2020; Ito et al., 2018; Li et al., 2022; Paul et al., 2019). In a visual-world context, cohort competition effects arise when listeners hear a target word that shares onset phonemes with one of the images on the screen; when hearing the onset of the target word (e.g., *beaker*), listeners fixate more on the image of a cohort competitor (e.g., *beetle*) than phonologically-unrelated distractors (e.g., *carriage*) (e.g., Allopenna et al., 1998). The onset of competition effects follows a similar time-course in both adults and children, though effects continue longer in young children (Sekerina & Brooks, 2007).

Experiment 1 used two different visual displays to see how each is affected by the noise introduced in web-based experimentation. Experiment 1A used a simple two-image display (with images on the left and right), similar to many infant preferential-looking studies. Experiment 1B used the four-image display that is common in visual-world studies (one image in each quadrant). Experiment 1B's four-image display further allows us to assess the performance of the eye-tracking methods on horizontal and vertical look discrimination.

The experiment methods and WebGazer phonemic cohort analysis were preregistered (https://osf.io/cn3ur/). The analysis of the webcam video data was exploratory. Prior to conducting Experiment 1, we ran a pilot experiment (N=24) to assess WebGazer's performance with adult participants (see Supplementary Materials).

Methods

A more detailed description of the methods is available in the Supplementary Materials. All experiments reported in this paper were approved by the Harvard University-Area Committee on the Use of Human Subjects.

Participants

Experiment 1 had 64 participants of five and six years of age who were native monolingual speakers of American English. Half completed Experiment 1A (N=32, 14 F, 18 M; M_{age}=5.8 years, SD=0.6, range=5;0–6;11), and half completed Experiment 1B (N=32, 20 F, 12 M; M_{age}=6.2 years, SD=0.5, range=5;0–6;11). Our sample size (32 participants per experiment) is similar to psycholinguistic experiments in general and to previous studies of the phonemic cohort effect (e.g., Farris-Trimble & McMurray, 2013; Huettig & McQueen, 2007). Informed written consent was received from the parent or guardian for their child's participation. Participants were compensated with a $5.00 gift card.

Materials

We selected 36 target–cohort pairs with onset overlap of one or more phonemes. As a control, each target word was pseudo-randomly assigned a competitor from another target–cohort pair with no onset overlap. The experiments consisted of 36 trials (one per word pair). The trial displays included a target image (corresponding to the target word) and a competitor image. In Experiment 1B, the displays also included two pseudo-randomly assigned distractor images whose names had different onsets from the target and competitor.[1] The trials were rotated through two conditions in two presentation lists. In the cohort condition, the competitor image depicted the cohort pair of the target (e.g., the target *milk* appeared with the competitor *mitten*). In the control condition, the target appeared with its control competitor (e.g., the target *milk* appeared with the competitor *windmill* from the cohort pair *window – windmill*). The cohort effect was assessed by comparing looks to the competitor images in the cohort and control conditions.

The experiments were built in PCIbex (Zehr & Schwarz, 2018) using PCIbex's implementation of WebGazer v2 and were completed in the participant's web-browser.

[1] One target (*doctor*) was accidentally assigned a distractor (*dolphin*) that shared onset sounds, so this trial was omitted from the Experiment 1B analysis.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

To accommodate the variability in screen-sizes across participant computers, stimulus size and location were defined by browser window size (equivalent to screen-size since the experiment was displayed fullscreen). Images appeared on canvases centered in their quadrant or half of the screen (Figure 1). Throughout each trial, WebGazer tracked looks to these canvases. When WebGazer detected a look to a canvas, the canvas border turned purple.[2]

Procedure

Participants completed the experiment while in a Zoom teleconference call with the experimenter(s), and the session was recorded via the Zoom meeting recording function. The participant opened the link to the experiment on their computer in Google Chrome or Mozilla Firefox and used the Zoom screen-sharing function to share the display with the experimenter. Participants using a non-Mac computer (with the exception of one Chromebook user) turned off their Zoom video prior to opening the experiment, as piloting revealed that many of these computers do not allow the same webcam to be used by Zoom and WebGazer simultaneously.

At the beginning of the experiment, the participant completed an audio check and a WebGazer calibration sequence. As we were interested in the range of calibration accuracy that would be obtained with our sample, we did not specify a minimum calibration threshold. After calibration, participants completed three practice trials followed by the 36 experimental trials. Each trial started with a calibration check. Next, the images appeared. After 2000ms, participants heard pre-recorded audio instructions

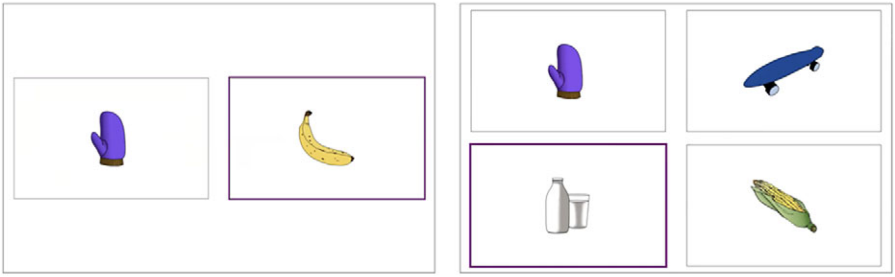

Figure 1. Example Experiment 1A (left) and Experiment 1B (right) trials. Each competitor image (e.g., *mitten*) appeared with its own target in the cohort condition (e.g., *milk*, right) and with another target in the control condition (e.g., *banana*, left). Image canvas borders turned from gray to purple when WebGazer estimated eye-gaze to fall on the image. Stills include images from Duñabeitia et al. (2018) and Rossion and Pourtois (2004).

[2] This color-change functionality allowed participants to use their eyes to select images from the screen (see Supplementary Materials for additional information about the task instructions given to participants). Initial piloting of web-based tasks with young children revealed that they were not familiar with how to use a computer mouse or trackpad, and click-based selection responses thus prompted a large number of participant looks directed at these tools instead of on the screen. Piloting with the color-change functionality indicated that it kept participants' attention on the screen, gave them a sense of agency in the task, and was not distracting.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

telling them to *Look at the + [target word].* The images remained on screen for 2250ms after audio offset. The full experiment session took approximately 20–30 minutes.

Analysis

The data for Experiments 1A and 1B were analyzed separately. All analyses were conducted using R v4.1.0 (R Core Team, 2021).

WebGazer

In each trial, WebGazer recorded looks from trial onset to two seconds after audio offset. In each sample, a 0 or 1 was recorded for each image canvas indicating whether or not participant gaze fell upon it (0=no, 1=yes). Sampling rate varied by participant, likely dependent upon their computer, webcam, and internet connection (grand mean time between samples=96ms, SD=43ms).[3] Samples which recorded no looks to any of the image canvases were excluded from analysis (41.24% of Experiment 1A samples; 27.07% of Experiment 1B samples). To regularize sampling rates prior to analysis, we analyzed gaze locations in bins of 100ms. A time bin received a value of 1 for a canvas if at least 50% of recorded looks within the bin fell on that canvas.

We preregistered a cluster permutation analysis to investigate competitor looks 0–2000ms after target onset (e.g., Hahn et al., 2015; Yacovone et al., 2021). This analysis assessed the effect of interest at each time step using generalized linear mixed-effect models (GLMMs) with a binomial distribution and logit link (step size=100ms).[4] All models in the present study were fit using the {lme4} package v1.1-27.1 (Bates et al., 2015). The models had looks to the competitor image (0, 1) as the dependent variable, a fixed effect of condition (cohort, control), and random slopes and intercepts for condition by participant and item. Item was individuated by competitor image identity to account for variance in properties of the competitor images. An effect was considered reliable at a step if the absolute value of its z-value was greater than 2 (Gelman & Hill, 2007).[5] A minimum

[3] Note that this average sampling rate (approximately 10 Hz) is slower than observed in some other WebGazer investigations, in which sampling rates range from 14–21 Hz (e.g., Prystauka et al., 2023; Semmelmann & Weigelt, 2018; Vos et al., 2022). Vos et al. (2022) and Prystauka et al. (2023) both implemented exclusion criteria to omit participants with a sampling rate below 5 Hz. Applying this same exclusion criteria to Experiment 1 (resulting in omission of n=2 participants), the mean time between samples is 93ms (SD=38ms), or approximately 11 Hz, suggesting that the slower sampling rate observed in Experiment 1 is not due to the lack of exclusion criteria but rather reflects the variability of web-based experimentation.

[4] It is important to note that while cluster-based permutation analyses provide information about the presence of effects, they cannot be used to make inferences about the onset and duration of these effects (for discussion, see Fields & Kuperberg, 2019; Groppe et al., 2011; Sassenhagen & Draschkow, 2019). As there are no corrections for multiplicity, false positives may emerge in the initial cluster identification, meaning that researchers cannot make inferences about effect significance at any one time bin in the cluster (including the first or final time bins). In addition, the cluster-mass permutation test does not assess how adding or removing time bins from the cluster (e.g., at the beginning or end) influences its overall reliability. Furthermore, cluster duration is sensitive to data quantity, power, and the chosen threshold for including time bins within a cluster, which could lead to under- or overestimations of the extent of effects.

[5] If a model failed to converge at a step (excluding singular fit warnings), we did not use the computed model estimates for that step. Instead, following Yacovone et al. (2021), we used the model estimates from the prior step; if the model at the first step did not converge, the z-value was set to zero. This procedure prevents

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

of two sequential reliable effects were required to comprise a cluster. To assess cluster reliability, we performed 1000 simulations reshuffling the condition labels for each participant. In each simulation, we summed the z-values of the adjacent steps in identified clusters to obtain a z-sum statistic. We compared the z-sum of the observed cluster to the distribution of each simulation's largest z-sum. A p-value for the observed cluster was determined by its position in this distribution (e.g., for a p-value of <0.05, 95% of the z-sums in the distribution must be greater than or equal to the observed statistic).[6]

We also analyzed the effect of condition on competitor looks in two time windows: 300–700ms after target onset (preregistered) and 600–1000ms after target onset (exploratory to account for a potential WebGazer delay in look detection). The results of these analyses are broadly consistent with the findings from the cluster analyses reported below and appear in the Supplementary Materials.

We conducted an additional exploratory analysis to investigate when target image looks were reliably different from chance in each condition. For each condition, we performed cluster permutation analyses assessing looks to the side of the screen containing the target image 0–2000ms after target onset; for Experiment 1B, we performed separate analyses for the horizontal and vertical side distinctions. In Experiment 1A, a look was considered to fall on the same side of the screen as the target if it fell on the target image; the analysis thus assesses the likelihood of target image looks. In Experiment 1B, a look was considered to fall on the same side of the screen as the target if it fell on the target or on the image vertically-adjacent (for the horizontal-side analysis) or horizontally-adjacent (for the vertical-side analysis). The analyses followed the same procedure described above, except that to assess reliability, we reshuffled the trial image location configurations by participant (thus preserving for each participant the overall number of target and non-target images appearing in each quadrant). The GLMMs computed at each step had target side looks (0, 1) as the dependent variable and random intercepts for participant and item (i.e., target image identity); as the model had no fixed effect, the likelihood of target side looks was compared to chance (50%). This analysis allows us to identify when each method is able to discriminate looks to the target quadrant along both the horizontal and vertical dimensions. For Experiment 1B, we supported the results of this analysis with a multinomial regression analysis assessing when looks differed between the target and the horizontally-, vertically-, and diagonally-adjacent images (see Supplementary Materials); the results align with the target side looks analyses.

Webcam video annotation

To gain further information about the eye-gaze patterns of our participants, we hand annotated gaze direction in the webcam videos of all participants who were able to keep their Zoom video on as they completed the experiment. Trial onsets times were identified from Zoom screen recordings using Python scripts that detected when the colored stimulus images appeared on screen (Anthony Yacovone, personal communication). These onsets were used to divide the continuous webcam videos into separate trial videos.

models that do not converge properly from breaking up or prematurely ending a cluster. There were no steps with non-convergence in the analyses of the observed data.

[6] In this analysis, it is possible to produce a p-value equal to zero if 0% of z-sums in the distribution of simulated statistics are greater than or equal to the observed statistic. We report these p-values as $p < 0.001$.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

Coders (blind to condition and target/competitor location) annotated gaze direction for each frame of these videos (annotation script by Anthony Yacovone).

Paralleling the WebGazer analysis, samples that were not coded as looks to one of the image locations were removed from analysis (i.e., center looks, blinks, etc.) (34.72% of Experiment 1A samples; 23.24% of Experiment 1B samples). The webcam videos had 40ms between samples. To compare to the WebGazer data, we analyzed gaze locations in bins of 100ms, following the binning procedure described above.

All videos were annotated by a single coder. To assess reliability, each video was additionally annotated by a secondary coder. Within our cluster analysis window (0–2000ms after target onset), inter-coder agreement was 92.18% in the Experiment 1A dataset and 90.02% in the Experiment 1B dataset (see Supplementary Materials for details). We performed the same analyses on the webcam video data as on the WebGazer data.

WebGazer results

Ten Experiment 1A trials across eight participants and 18 Experiment 1B trials across seven participants were omitted from the WebGazer analysis because no data were saved for them on our server.

Calibration scores

Participant calibration scores in the initial calibration sequence ranged from 2–80% across Experiments 1A and 1B, with an average of 43% (SD=18, see Supplementary Materials for plots and more detail). Mean participant calibration scores during the calibration checks at the beginning of each experimental trial ranged from 8–50%, with an average of 30% (SD=11).

Experiment 1A

Figure 2 illustrates the increase in looks to the target image in the WebGazer output following target word articulation in both the cohort and control conditions. This pattern was similar for targets on the left and right of the screen (see Supplementary Materials). While there was a substantial rise in target looks in both conditions (~75% of looks), this rise was smaller than commonly observed in two-image studies with children and adults (e.g., 80–85% with adults and three to four year-olds in Simmons, 2017).

Target looks were reliably different from chance in clusters starting 800ms after target onset in the control condition (z-sum=64.94, p<0.001) and 1000ms after target onset in the cohort condition (z-sum=61.82, p<0.001).

Figure 3 focuses on the cohort effect by plotting looks to the competitor image in the cohort and control conditions. Prior to target word onset, looks to the competitor image were at chance (50%). These looks began to decline approximately 700ms after target word offset (as target looks increased). Our analyses explored whether this decline was faster in the control condition than the cohort condition. The analysis identified a reliable difference in competitor looks between conditions in a cluster 900–1099ms after target onset (z-sum=4.87, p=0.02).

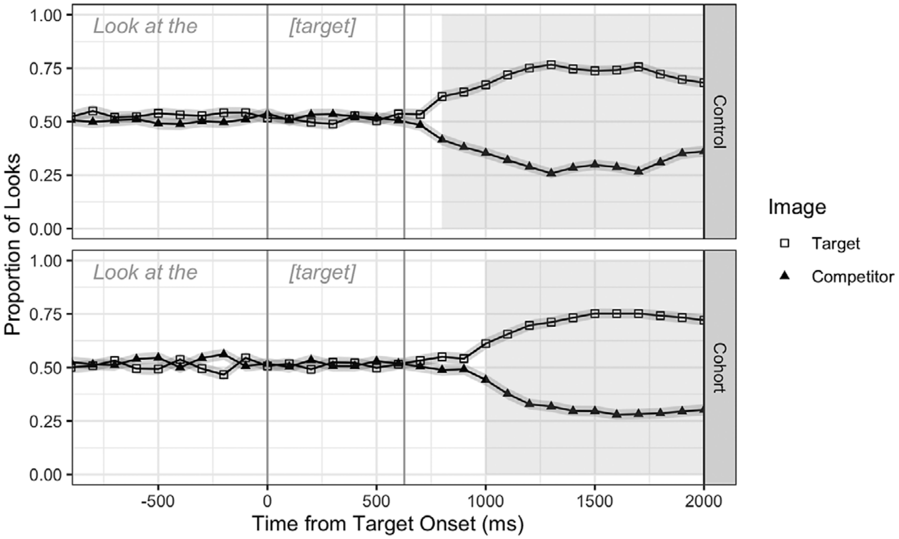

Figure 2. Mean WebGazer looks to the target and competitor images by condition in Experiment 1A. Ribbons indicate standard error. Vertical lines indicate average target word duration. Shading indicates when looks to the target image differed from chance.

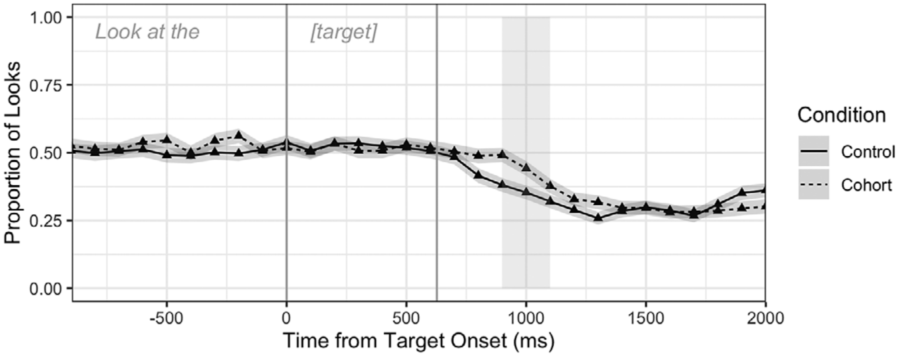

Figure 3. Mean WebGazer looks to the competitor image by condition in Experiment 1A. Ribbons indicate standard error. Vertical lines indicate average target word duration. Shading indicates when looks between conditions were reliably different in the cluster analysis.

Experiment 1B

Figure 4 shows looks to the target image, competitor image, and two distractor images (collapsed) as detected by WebGazer in the cohort and control conditions. In both conditions, WebGazer detected increased looks to the target image following target word onset. However, the effects appeared smaller than in previous studies (≤50% in the present study vs. >60% with five and six year-olds in Sekerina & Brooks, 2007).

In the control condition, looks to the side of the screen containing the target were reliably different from chance in clusters starting 900ms after target onset along the horizontal axis (z-sum=62.73, p<0.001) and 1200ms after target onset along the vertical

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

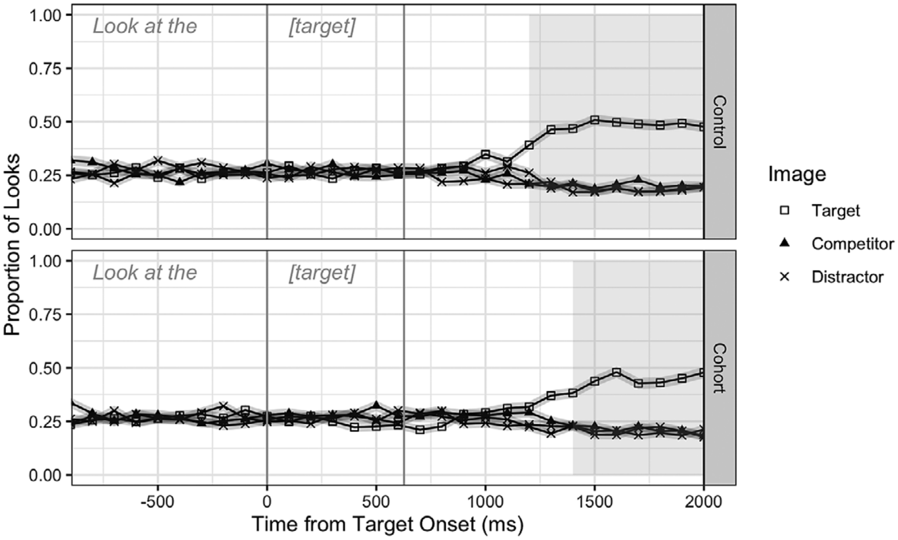

Figure 4. Mean WebGazer looks to the target image, competitor image, and distractor images (collapsed) by condition in Experiment 1B. Ribbons indicate standard error. Vertical lines indicate average target word duration. Shading indicates the temporal overlap of the clusters when target side looks differed from chance in both the horizontal and vertical directions.

axis (z-sum=29.10, p<0.001). In the cohort condition, clusters emerged 1000ms after target onset for the horizontal-side distinction (z-sum=50.49, p<0.001) and 1400ms after target onset for the vertical-side distinction (z-sum=17.83, p=001).

The observed clusters for the horizontal-side distinction had similar onsets to those in Experiment 1A (800ms in the control condition, 1000ms in the cohort condition) – however, the observed clusters for the vertical-side distinction started 300–400ms later, suggesting that WebGazer may have more difficulty discriminating looks along the vertical axis. Figure 5 plots participant looks to the target and distractor (non-target) images in the control condition 1200–2000ms after target onset (when participants were likely fixating on the target quadrant, according to WebGazer). In this window, there were more looks to the vertical distractor than the other non-target images, supporting the hypothesis that WebGazer has increased difficulty discriminating vertical looks (this pattern was confirmed in an exploratory multinomial analysis; see Supplementary Materials). A figure showing target and distractor looks by target location is available in the Supplementary Materials.

Figure 6 plots looks to the competitor image in the cohort and control conditions. Prior to target word onset, looks to the competitor image were at chance (25%). These looks began to decrease approximately 1200ms after target onset. The cluster analysis did not identify any clusters where competitor looks differed in the two conditions. Thus, we did not replicate the phonemic cohort effect.

Webcam video annotation results

We had video data for 13 of 32 participants for each experiment.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

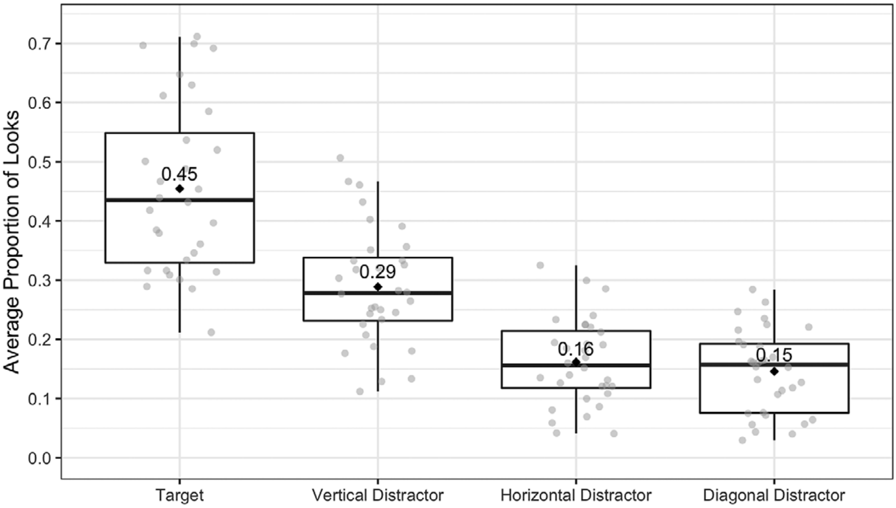

Figure 5. Boxplot of participant WebGazer fixation proportions to the target and non-target images in the Experiment 1B control trials from 1200–2000ms after target onset. Mean fixation proportions for each image are labeled and identified by black diamonds. The gray points represent participant means.

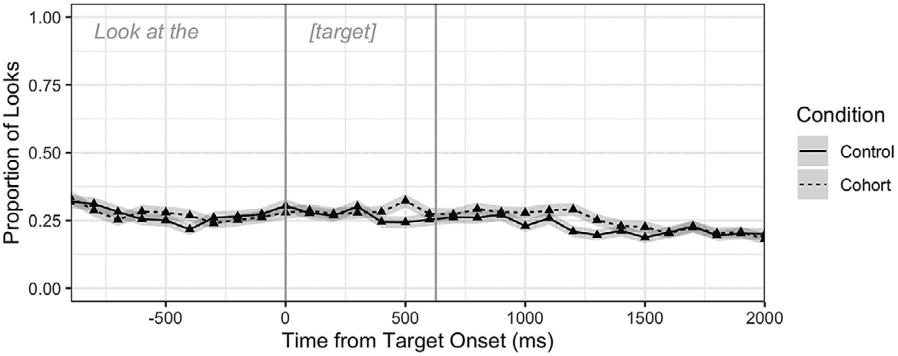

Figure 6. Mean WebGazer looks to the competitor image by condition in Experiment 1B. Ribbons indicate standard error. Vertical lines indicate average target word duration.

Experiment 1A

Figure 7 plots looks to the target and competitor images in the cohort and control conditions as detected by hand annotation and WebGazer for the 13 participants with video data. Webcam video annotation identified a higher proportion of target image looks than WebGazer. The pattern of performance was similar for targets on the left and right of the screen (see Supplementary Materials).

Target looks were reliably different from chance in clusters starting 500ms after target onset in the control condition (z-sum=93.02, p<0.001) and 800ms after target onset in the cohort condition (z-sum=75.36, p<0.001). These clusters started earlier than in the WebGazer data from the same participants, in which the corresponding clusters began

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

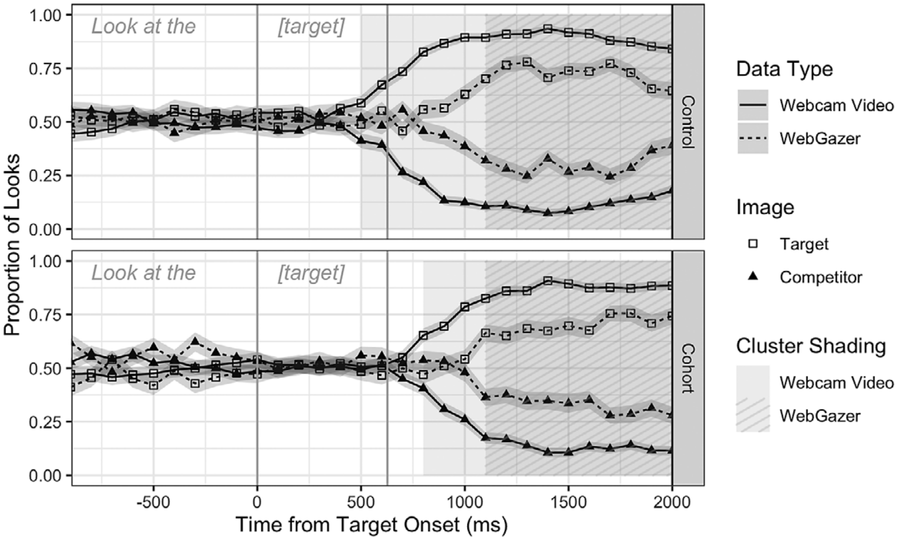

Figure 7. Mean looks to the target and competitor images by condition in the Experiment 1A annotated webcam video data and in the WebGazer data from the same participants. Ribbons indicate standard error. Vertical lines indicate average target word duration. Shading indicates when looks to the target image differed from chance.

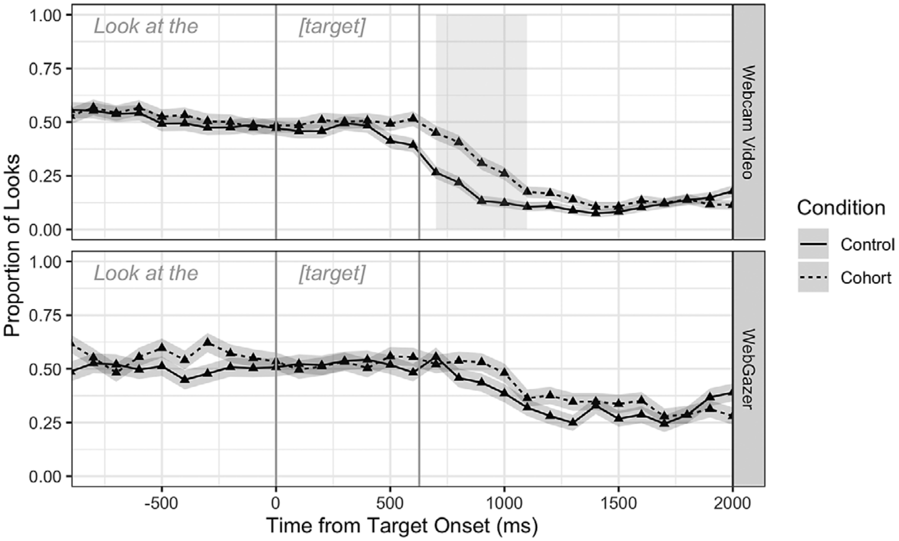

Figure 8. Mean looks to the competitor image by condition in the Experiment 1A annotated webcam video data and in the WebGazer data from the same participants. Ribbons indicate standard error. Vertical lines indicate average target word duration. Shading indicates when looks between conditions reliably differed.

1100ms after target onset in both the control (z-sum=38.65, p<0.001) and cohort (z-sum=35.28, p<0.001) conditions.

Figure 8 shows looks to the competitor image in the cohort and control conditions. In the video data, competitor looks in the control condition decreased during target word

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

articulation, whereas looks in the cohort condition did not decrease until target word offset. In contrast, in the WebGazer data from the same participants, competitor looks decreased only after target word offset in both conditions (similar to the pattern observed in the full WebGazer dataset), and competitor looks were more similar in the two conditions. In the video data, the analysis identified a reliable difference in competitor looks between conditions in a cluster 700–1099ms after target onset (z-sum=13.26, p=0.001), thereby showing evidence of a phonemic cohort effect. A cluster analysis of the corresponding WebGazer data did not identify any clusters.

Experiment 1B

Figure 9 plots looks to the target and competitor images in the cohort and control conditions, as identified by webcam video annotation and WebGazer for the same 13 participants (plots including distractor images are available in the Supplementary Materials). Target looks rose earlier and reached higher proportions in the video data than the WebGazer data.

In the video data for the control condition, looks to the side of the screen containing the target were reliably different from chance in clusters starting 600ms after target onset along the horizontal axis (z-sum=60.27, p<0.001) and 700ms after target onset along the vertical axis (z-sum=63.13, p<0.001). In the cohort condition, clusters emerged 800ms after target onset for the horizontal-side distinction (z-sum=65.33, p<0.001) and 600ms after target onset for the vertical-side distinction (z-sum=75.98, p<0.001).

In the WebGazer data from the same participants, the detection of target looks appeared considerably later. In the control condition, target-side looks were reliably

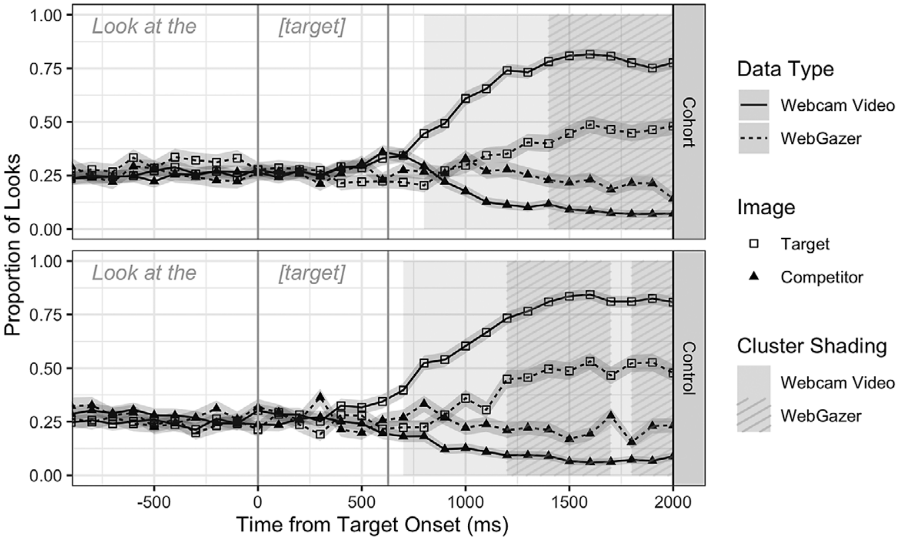

Figure 9. Mean looks to the target and competitor images by condition in the Experiment 1B annotated webcam video data and in the WebGazer data from the same participants. Ribbons indicate standard error. Vertical lines indicate average target word duration. Shading indicates the temporal overlap of the clusters when target side looks differed from chance in both the horizontal and vertical directions.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

different from chance in clusters emerging 1200ms after target onset along both the horizontal (z-sum=34.73, p<0.001) and vertical (z-sum=15.96, p=0.001) axes.[7] In the cohort condition, clusters emerged 1000ms after target onset for the horizontal-side distinction (z-sum=36.28, p<0.001) and 1400ms after target onset for the vertical-side distinction (z-sum=15.80, p=0.001).

Figure 10 shows participants' looks to the target and distractor images in the control condition 700–2000ms after target onset (when participants were likely fixating on the target quadrant, according to the video annotation) for the webcam video data. The proportion of looks to the target was higher than during detected target fixations in the full WebGazer sample (Figure 5), and there were fewer distractor looks. Similar to the full WebGazer sample, there was a slight preference for vertical distractors over the other non-target images (this pattern was confirmed in an exploratory multinomial analysis; see Supplementary Materials) – however, the relative differences were smaller in the webcam video data. A figure showing target and distractor looks by target location is available in the Supplementary Materials.

Figure 11 shows looks to the competitor image in the cohort and control conditions. In the video data, looks to the competitor image in the cohort condition increased during target articulation, while looks in the control condition decreased. In the WebGazer data from the same participants, there was no obvious difference between conditions. In the video data, the analysis identified a reliable difference in competitor image looks between conditions in a cluster 600–999ms after target onset (z-sum=11.19, p<0.01), thus finding evidence of a phonemic cohort effect. A cluster analysis of the corresponding WebGazer data did not identify any clusters.

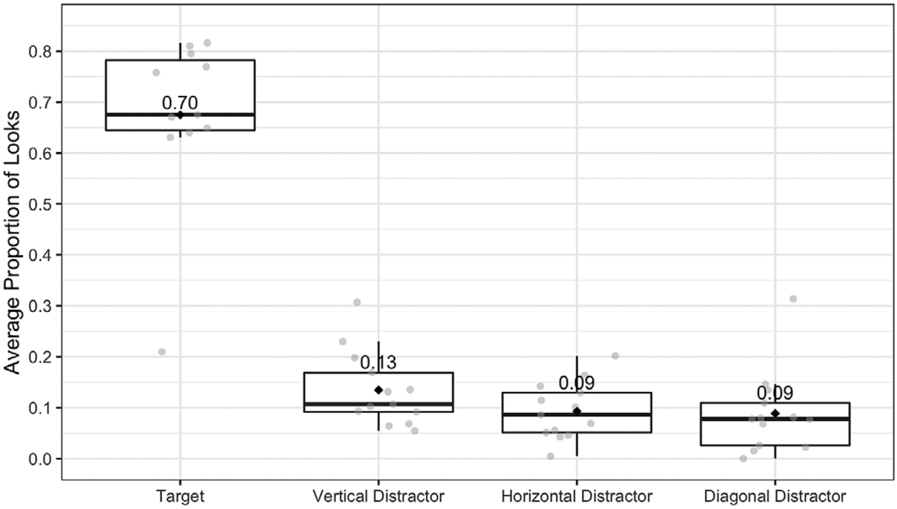

Figure 10. Boxplot of participant fixation proportions to the target and non-target images in the Experiment 1B control trials from 700–2000ms after target onset for the annotated webcam video data. Mean fixation proportions for each image are labeled and identified by black diamonds. The gray points represent participant means.

[7] The analysis of vertical-side looks identified two clusters: one 1200–1699ms after target onset (z-sum=15.96, p=0.001) and one 1800–1999ms after target onset (z-sum=6.65, p=0.049). The effect in the 1700ms bin had a z-score of 1.98, so it did not meet the threshold to be included in a cluster.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

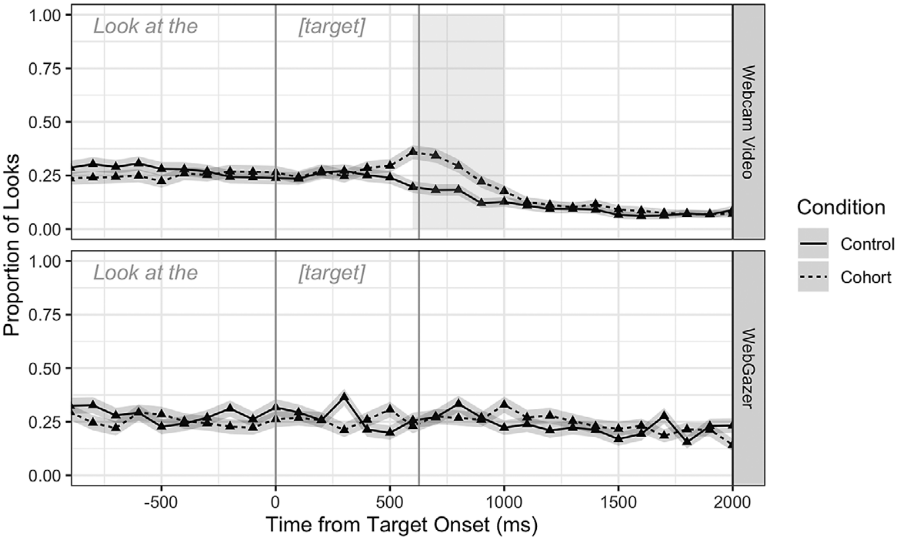

Figure 11. Mean looks to the competitor image by condition in the Experiment 1B annotated webcam video data and in the WebGazer data from the same participants. Ribbons indicate standard error. Vertical lines indicate average target word duration. Shading indicates when looks between conditions reliably differed.

Experiment 1 summary

Experiment 1 used a standard visual-world task to assess the relative performance of two webcam-based eye-tracking methods with five to six year-old children: automatic WebGazer gaze coding and hand annotation of gaze direction from recorded webcam videos. Both methods detected increased looks to named (target) images in both two- and four-image displays. However, the rise in target fixations was lower and later in the WebGazer data compared to in-lab experiments with children of the same age or younger (e.g., Sekerina & Brooks, 2007; Simmons, 2017). The annotated video data, on the other hand, looked more like data collected in in-lab experiments: the onset of target looks was faster, and the proportion of target looks was considerably higher than in simultaneously-collected WebGazer data. Interestingly, for both methods, unrelated images vertically-adjacent to the target received more looks than distractor images in the other locations of the display; this pattern was especially notable in the WebGazer data.

The differences between the two methods were particularly pronounced in the analysis of the phonemic cohort effect. In the video data, the cohort effect emerged in both the four- and two-image displays in clusters beginning 600–700ms after target onset and was detectable in a sample of just 13 children. This effect is later than observed in previous lab-based studies, in which cohort effects began 200–400ms after target onset (e.g., Allopenna et al., 1998; Huettig & McQueen, 2007; Sekerina & Brooks, 2007). While this difference could reflect our small sample size or a difference in our analysis method, it is consistent with other research using webcam video annotation (i.e., the web-based replication of Allopenna et al., 1998 by Ovans, 2022). In the WebGazer data, the effect was detectable only in the two-image display with a larger sample (N=32), and this effect window emerged later (900ms after target onset). These results suggest that while WebGazer can detect robust fixation patterns like target looks, webcam video annotation is better suited to detecting more fine-grained effects.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

In Experiment 1 we tracked looks in a binary fashion, monitoring whether or not a look fell inside a particular region. While this measure reflects how visual-world studies are generally conducted, we cannot tell from these results how close WebGazer's gaze estimates are to the true locations of visual stimuli. Experiment 2 explores WebGazer's accuracy more directly. This additionally allows us to address one limitation of Experiment 1: because our image canvases did not cover the full halves (Experiment 1A) or quadrants (Experiment 1B) of the screen, gazes that were estimated to fall near a canvas, but not within it, may have been coded as looks in our video data but not in the WebGazer data.

Experiment 2: fixation task

Experiment 2 used a visual-fixation task to investigate the spatial and temporal resolution of WebGazer's gaze estimation with four to twelve year-old children. This task was adapted from Slim and Hartsuiker (2022) ("S&H2022"). The experiment had four goals: i) to assess the feasibility of conducting web-based eye-tracking tasks with children without an experimenter present; ii) to assess how closely WebGazer estimates correspond to stimulus locations; iii) to assess whether there are age-related differences in WebGazer performance between four and twelve years; and iv) to assess whether the accuracy of quadrant-based analyses with WebGazer is improved by using larger canvases.

Participants

The study included 45 participants between four and twelve years of age (Table 1). Participants spoke American or British English natively. Participants were not required to be monolingual, as the experiment was non-linguistic. Three participants in Experiment 2 previously took part in Experiment 1 during a different experiment session. Informed written consent was received from the parent or guardian for their child's participation; child participants additionally provided written assent. Participants were compensated with a $5.00 gift card.

Materials

The experiment was built in PCIbex (Zehr & Schwarz, 2018) using WebGazer v2 and was completed in the participant's web-browser. The stimuli were modeled on those from S&H2022. Participants looked to fixation crosses that appeared in 13 possible screen positions (Figure 12). Each fixation cross appeared in each location six times, resulting in 78 total trials. Trial order was randomized for each participant. To accommodate

Table 1. Experiment 2 participant ages

Age group	Count	Mean age (months)	Range (years;months)
4–5 years	12 (7 F, 5 M)	62.9 (SD=5.2)	4;6–5;9
6–7 years	12 (5 F, 7 M)	81.8 (SD=6.6)	6;0–7;7
8–9 years	11 (4 F, 6 M, 1 NB)	104.4 (SD=4.5)	8;3–9;6
10–12 years	10 (7 F, 3 M)	132.5 (SD=9.6)	10;0–12;8

variability in computer screen-sizes, the experiment was completed in fullscreen, and stimulus size and location were defined by browser window size. To make the task more fun for child participants, the 78 trials were divided into six blocks: in each block, the fixation cross appeared in a different color and was accompanied by a different audio sound effect.

Procedure

The experiment was completed by participants from their own computers, unsupervised by researchers. An experiment access link was sent to the parent's email. Participants were asked to complete the experiment on a computer or laptop using either Google Chrome or Mozilla Firefox. An adult was asked to help the child get set up and to remain in the room as they completed the task.

The experiment started with an introductory sequence that walked participants through an audio check, the WebGazer calibration, and the experiment instructions. Following the audio check, the sequence included both written and auditory instructions so that it would be accessible to both child participants and adult supervisors. As in Experiment 1, we did not specify a minimum calibration threshold. Participants were instructed to look at the plus signs that appeared on the screen; they were instructed to look at them as fast as they could and to stare at them until they disappeared.

To start each block of the task, the participant pressed the spacebar, which initiated a calibration check (resulting in seven total calibration scores per participant). The trial structure was the same as in S&H2022. Each trial began with a small black fixation cross (+) appearing in the center of the screen for 500ms (font size defined as 5% of the screen height). This cross then disappeared and the colored target fixation cross appeared on screen for 1500ms (size defined as 10% of the screen height). The trial then ended, and the

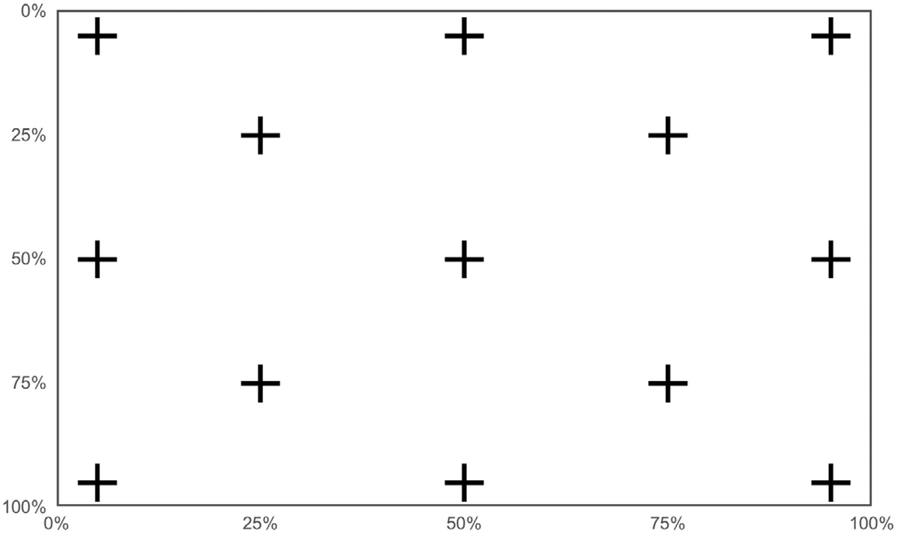

Figure 12. The 13 possible target stimulus locations in Experiment 2. The panel represents the full experiment screen (the axis labels indicate percentage of screen-size).

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

next trial began automatically. The experiment took approximately 10–15 minutes to complete.

Data processing

WebGazer tracked participants' eye-movements from target stimulus onset to trial offset. We recorded looks to canvases covering each quadrant of the screen as a binary variable. These canvases together covered the entire screen (each 50% of browser window height and width). We also recorded coordinate estimates of gaze location (in pixels). If either the x- or y-coordinate estimate was missing in the recorded WebGazer data, the sample was omitted from the data prior to processing (0.53% of samples).

Data processing followed the procedure outlined by S&H2022 using the scripts made available in their OSF repository (https://osf.io/yfxmw/). We aggregated the data into 100ms bins, calculating for each bin the mean x- and y-coordinate estimates and mean looks to each quadrant canvas (quadrant looks were later binarized for analysis). We restricted the dataset to bins ranging from 0–1500ms after trial onset, resulting in exclusion of 170 out of 3484 recorded bins (4.88%). To account for participants' different screen-sizes, we converted the pixel coordinate estimates to a distance metric based on screen-size proportion, such that the pixel in the center of the screen had coordinates (0.5, 0.5) and the pixel in the bottom right corner had coordinates (1,1). For each bin, we calculated the Euclidean distance between the estimated gaze location and the center of the target fixation cross (in proportion of screen-size) using the formula below.

$$\sqrt{\left(x_{target} - x_{gaze\ estimation}\right)^2 + \left(y_{target} - y_{gaze\ estimation}\right)^2}$$

In some of our analyses, we compare the Experiment 2 data to S&H2022's adult data accessed from their OSF repository.

Results

26 trials across 10 participants were omitted from the analysis because no WebGazer data were saved for them on our server.

Calibration scores

Participant calibration scores in the initial calibration sequence ranged from 6–80%, with an average score of 52% (SD=16). The mean participant scores for the six calibration checks ranged from 4–67%, with an average of 39% (SD=13). Table 2 summarizes participant mean calibration scores (calculated using all seven scores for each participant) by age group.

As in S&H2022, participant mean calibration scores were significantly correlated with webcam sampling rates (measured in frames per second) (ρ=0.33, p=0.03), suggesting that WebGazer's estimates are more precise when there are more recorded samples. In addition, mean calibration score was significantly correlated with participant age in months (to one decimal place) (ρ=0.52, p<0.001), indicating that older participants tended to have higher calibration scores. This trend still holds when accounting for sampling rate (see Supplementary Materials for more details and plots).

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

Table 2. Experiment 2 mean participant calibration scores by age group

Age group	Mean score (%)	Range (%)
4–5 years	33 (SD=13)	11–50
6–7 years	40 (SD=12)	4–52
8–9 years	43 (SD=9)	28–52
10–12 years	51 (SD=10)	35–66

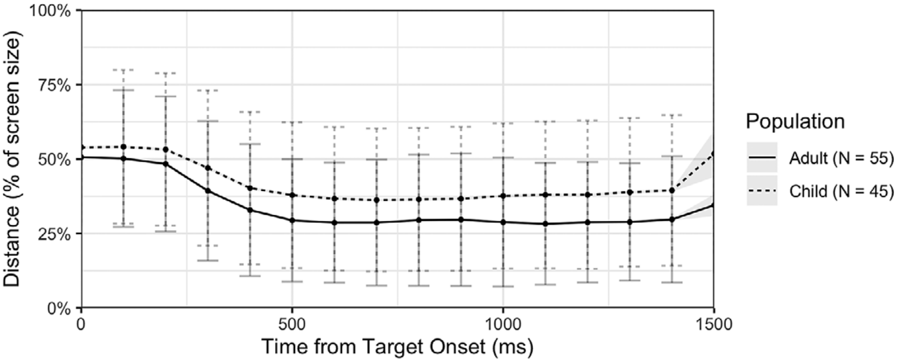

Figure 13. Mean Euclidean distance (in percentage of screen-size) from the target stimulus over the course of the trial. Error bars indicate standard deviation. Ribbons indicate standard error.

Euclidean distance from the target over time

To assess how closely WebGazer estimates match stimulus location, we plotted the mean Euclidean distance (in percentage of screen-size) between the target stimulus and estimated gaze location from stimulus onset to trial offset (Figure 13). The plot includes data for the Experiment 2 child participants as well as S&H2022's adult participants. In both populations, distance from the target began to decrease 200ms after stimulus onset and plateaued around 500ms after onset. While this timing is similar for the two populations, the Euclidean offset was larger and more variable for children, settling at an offset of approximately 38% of screen distance from the target.

To better understand the factors influencing this offset, we plotted mean distance over time broken down by calibration score (Figure 14) and child age group (Figure 15). [8] Figure 14 illustrates the relationship between mean calibration score and WebGazer's spatiotemporal accuracy: mean Euclidean offset was smaller for participants in higher calibration bins. Figure 15 suggests that there was also a relationship between Euclidean distance and age: offsets plateaued at the shortest distance for the 10–12 year-old participants, followed by the 8–9 year-old and 6–7 year-old participants, with the longest distance offsets for the 4–5 year-old participants. However, as discussed above, mean calibration score and age were correlated; therefore, it is not obvious from Figure 15 the

[8] Given variations in WebGazer sampling, there were fewer samples towards the end of the trial (see also S&H2022). We thus ended the plots at 1200ms, the latest time point for which we had enough samples to calculate standard errors in all bins for both plots.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

extent to which age contributed to Euclidean offset independently from calibration score. We address this below.

Looks in the fixation window

As in S&H2022, we analyzed a fixation time window 500–1500ms after target onset to assess WebGazer's spatial resolution when gaze had settled on the target location. Figure 16 plots the density of looks on the screen during this time window for all 13 target locations. Density plots of the quadrant fixations for the youngest and oldest age groups in our sample are available in the Supplementary Materials.

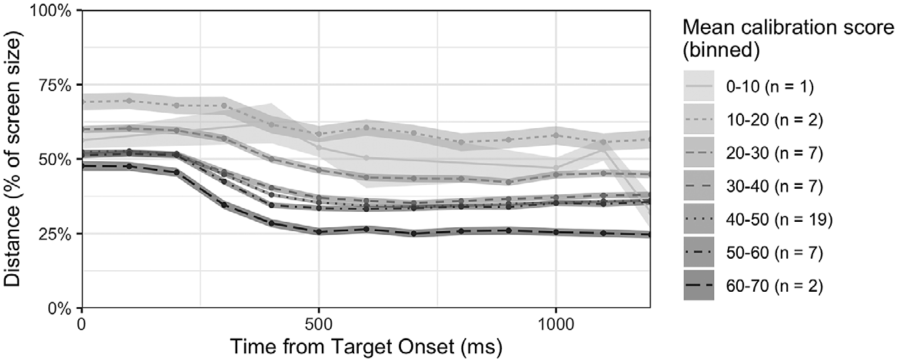

Figure 14. Mean Euclidean distance (in percentage of screen-size) from the target stimulus over the course of the trial, broken down by participant calibration score. Ribbons indicate standard error.

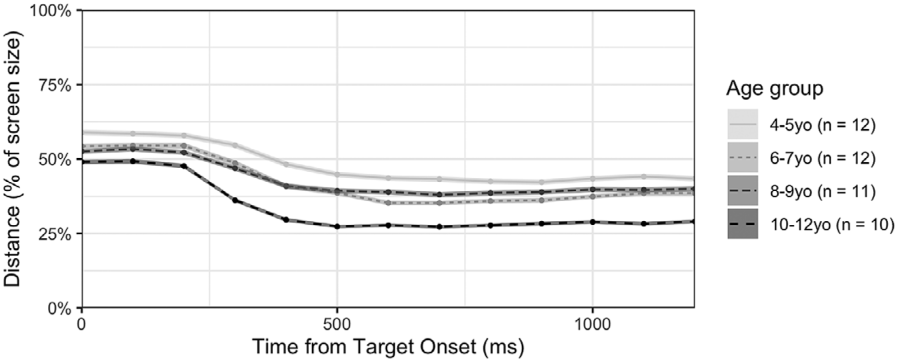

Figure 15. Mean Euclidean distance (in percentage of screen-size) from the target stimulus over the course of the trial, broken down by participant age bin. Ribbons indicate standard error.

For each location, estimated looks tended to fall around the stimulus, though the range in which the looks fell was large. In the plots for the targets appearing in the center of each quadrant (the second and fourth row of Figure 16), estimated looks often extended into quadrants other than the one containing the target stimulus, with particular overlap in the

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

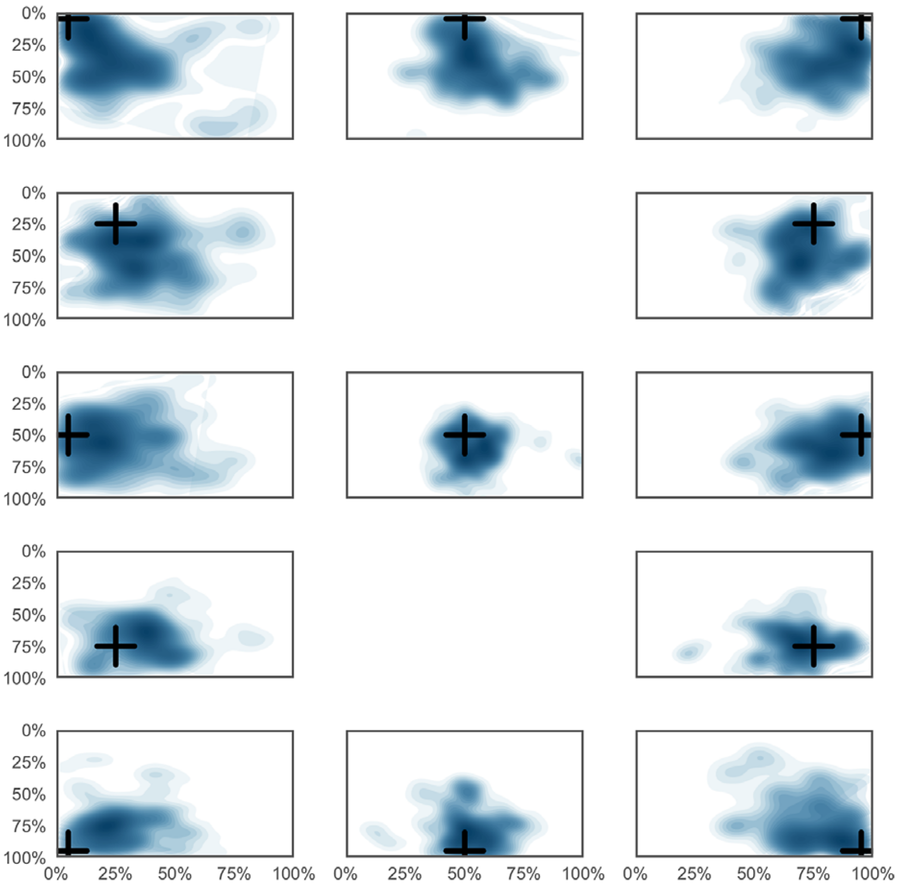

Figure 16. Density plots indicating estimated looks on the screen 500–1500ms after target onset for each possible target location. Each panel represents the full experiment screen (the axis labels indicate percentage of screen-size), and the black crosses indicate the center of the target locations.

vertical direction. In fact, within the fixation window, participants' mean vertical offsets between their estimated gaze location and the true stimulus location (M=0.27, SD=0.08) were greater than their mean horizontal offsets (M=0.21, SD=0.08) (t(44)=5.55, p<0.0001). WebGazer's reduced vertical accuracy appears particularly pronounced in the upper quadrants of the screen (the second row of Figure 16).

To assess the relative contributions of calibration accuracy and participant age to Euclidean distance offset during target fixations, we calculated the mean distance from the target during the 500–1500ms fixation window for each participant and computed a linear regression with fixed effects of mean calibration score and age in months (to one decimal place). Both the effects of mean calibration score (β=-0.004, t=-3.93, p<0.001) and age (β=-0.001, t=-2.30, p=0.03) were reliable.[9] Model comparison using ANOVA

[9] Given the correlation between the two model predictors, multicollinearity in the model was assessed by calculating the Variance Inflation Factor (VIF); VIF for both predictors was 1.38.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

revealed significant differences between models with both predictors and models with only calibration score (F(1,42)=5.27, p=0.03) and only age (F(1,42)=15.4, p<0.001). These results suggest that there was an effect of age on Euclidean offset that was distinct from the effect of calibration score.

Quadrant looks over time

In addition to investigating Euclidean distance over time, we also analyzed quadrant looks over time, allowing us to assess WebGazer's accuracy discriminating quadrant looks when using larger canvases than in Experiment 1B. We restricted the data to the trials in which the fixation cross appeared in the center of each screen quadrant. We binarized quadrant looks using the same procedure as in Experiment 1B. For comparison, we also binarized S&H2022's adult data in the same fashion.

Figure 17 plots looks to the target quadrant over time compared to the other quadrant locations (horizontally, vertically, or diagonally across from the target) for both populations. A plot showing quadrant looks by target location for the child participants is available in the Supplementary Materials. The pattern of target quadrant looks in the child data resembles that observed in Experiment 1B: looks to all quadrants began at chance (25%), and then looks to the target increased and plateaued around 50%.

To assess target quadrant looks, we performed the same target side analyses as we conducted for Experiment 1B. We analyzed looks from 0–1400ms after target onset given the reduced number of samples at the end of the trial. The analyses followed the same procedure as the Experiment 1B analyses, except quadrants were used instead of images, and item was defined as target location (top left, top right, bottom left, bottom right). In the Experiment 2 child data, looks to the side of the screen containing the target were

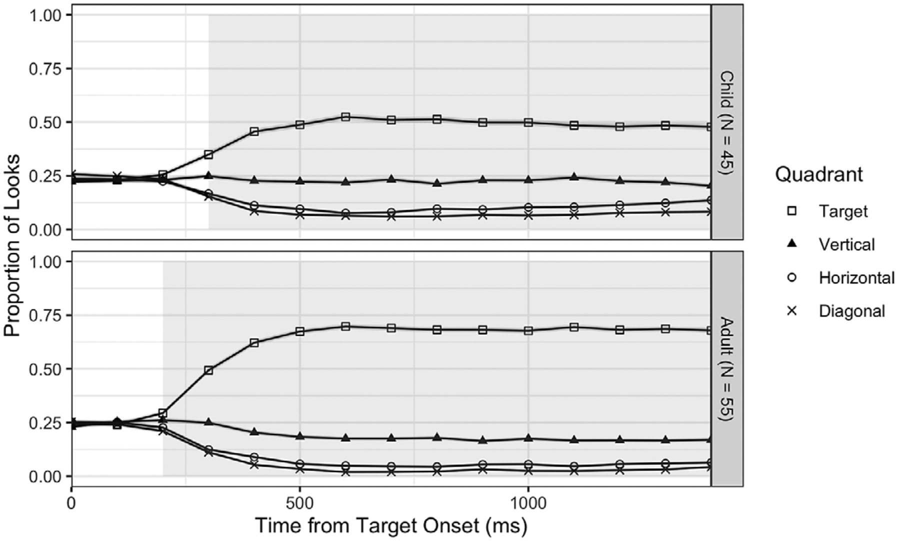

Figure 17. Quadrant looks over time for the Experiment 2 child participants and Slim and Hartsuiker's (2022) adult participants. Ribbons indicate standard error. Shading indicates the temporal overlap of the clusters when target side looks differed from chance in both the horizontal and vertical directions.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

reliably different from chance in clusters starting 300ms after target onset along both the horizontal (z-sum=92.71, p<0.001) and vertical (z-sum=56.30, p<0.001) axes (in the exploratory multinomial analysis, target quadrant looks similarly differed from looks to all other quadrants in a cluster starting 300ms after onset; see Supplementary Materials). In the S&H2022 adult data, clusters started at target onset for the horizontal-side distinction (z-sum=136.59, p<0.001)[10] and 200ms after target onset for the vertical-side distinction (z-sum=90.31, p<0.001), suggesting that WebGazer detected target quadrant fixations starting 200ms after target onset (in the exploratory multinomial analysis, looks to the target differed in a cluster starting 300ms after onset).[11]

In both the adult and child data, looks to the quadrant vertically adjacent to the target remained elevated compared to the other non-target quadrants during target fixations (Figure 17). This pattern was confirmed in an exploratory multinomial analysis (see Supplementary Materials).

Experiment 2 summary

The Experiment 2 results demonstrate that it is possible to conduct unsupervised web-based eye-tracking tasks with school-aged children. There was a sharp increase in looks toward the target shortly after it appeared, indicating that participants were able to perform the task without an experimenter to guide them. Furthermore, the data suggest that parents, acting on their own, were just as effective in setting up the experiment as parents guided by researchers; there was no significant difference between the mean calibration scores of participants in Experiment 1 and those of the same age range (five–six years) in Experiment 2 (t(7)=-0.74, p=0.49).

Experiment 2 assessed how closely WebGazer estimates track with stimulus location. The Euclidean offset between estimated gaze and target stimulus location was approximately 38% of screen-size. This offset is greater than observed in S&H2022's adult data (30% of screen-size) and much larger than reported for in-lab eye-trackers (see *General Discussion*). In addition, we found age-related differences in performance: calibration scores tended to be higher for older participants, and there was a relationship between participant age and Euclidean offset above and beyond the effect of calibration.

Analyzing the Experiment 2 data using the quadrant-based approach common for visual-world studies showed a similar overall pattern to Experiment 1B, suggesting that increasing the size of the tracked quadrant canvases did not substantially improve WebGazer data quality. Target quadrant looks increased faster in Experiment 2 than

[10] The early cluster onset identified in the horizontal target-side looks analysis appears to be driven by a slight preference for target-side looks in the 0ms and 100ms time bins (in both bins, 54% of recorded looks fell on the target side). The beta coefficients for these time bins (0.17 and 0.12, respectively) suggest that the effect was small; for reference, the beta coefficient 500ms after target onset (once target looks plateau in Figures 13 & 15) was >2.

[11] This timing differs from that identified in S&H2022's analysis, which identified clusters in the 0–200ms and 400–1400ms bins (Slim & Hartsuiker, 2022). Their analysis collapsed all non-target quadrants into a single other quadrant variable that received a 1 if there was a look to any quadrant other than the target quadrant; they then analyzed whether looks (0,1) differed based on focus (target quadrant, other quadrant). We did not perform our analyses this way due to concerns about dependencies in the data structure. Repeating our cluster analysis using this structure, we obtained the same two clusters identified by S&H2022 (in the bins 0–200ms and 400–1400ms after target onset). Using S&H2022's analysis structure on the Experiment 2 child data yields clusters in the 0–300ms and 500–1400ms bins.

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

Experiment 1, likely due to the larger quadrant sizes and the fact that attention was directed to the stimulus by a single visual cue (the target was the only item on screen), whereas in Experiment 1 participants needed to process linguistic input to determine which of multiple visual stimuli to fixate upon. Our results furthermore support the finding from Experiment 1 that WebGazer is less accurate at detecting vertical distinctions: during target fixations, offsets between estimated gaze locations and the true stimulus location were greater in the vertical direction than the horizontal direction, and in the quadrant-based analysis, there were elevated looks to the quadrant vertically adjacent to the target. This inaccuracy appears to be particularly pronounced in the top–down direction, with greater vertical offsets for targets appearing on the top half of the screen (Figure 16).

General discussion

The present study investigated the suitability of two webcam eye-tracking methods for child language research: automatic WebGazer gaze estimation and frame-by-frame annotation of gaze direction from webcam videos. Experiment 1 compared these two methods with five and six year-olds in a visual-world task replicating the phonemic cohort effect. The experiment used two display types: a two-image display with one image on each side of the screen (Experiment 1A) and a four-image display with one image in each quadrant (Experiment 1B). Experiment 2 investigated WebGazer's gaze estimation accuracy in an unsupervised visual-fixation task with four to twelve year-old children. Our results suggest that while it is possible to conduct webcam eye-tracking studies with children (supervised and unsupervised), the two eye-tracking methods differ in their spatiotemporal resolution and thus are not equally suitable for detecting all types of eye-movement patterns. In this Discussion, we discuss the spatiotemporal accuracy of the two methods, their ability to detect fine-grained linguistic effects, and recommendations for researchers conducting web-based eye-tracking experiments with children.

Spatiotemporal accuracy of the eye-tracking methods

Spatial resolution

Both webcam eye-tracking methods were sufficiently accurate to detect the preference to look at a target that either is explicitly mentioned (Experiment 1) or suddenly appears on the screen (Experiment 2). This is true both when target and foil occupy different halves of the screen (Experiment 1A) and when the target occupies one quadrant (Experiment 1B, Experiment 2). Nevertheless, webcam video annotation had a higher signal-to-noise ratio, as evidenced by a higher proportion of target looks than in simultaneously-collected WebGazer data (89% vs. 72% in Experiment 1A; 80% vs. 47% in Experiment 1B). In fact, the target looks in the video data parallel those from prior in-lab experiments using commercial eye-trackers with children of this age (Sekerina & Brooks, 2007; Simmons, 2017).

Experiment 2 confirmed WebGazer's reduced spatial accuracy compared to in-lab eye-tracking using a more fine-grained distance metric. Target fixations as detected by WebGazer were approximately 38% of the screen distance from the true stimulus location, compared to an offset of 1–2% (0.4–0.9° of the visual angle) reported for Tobii TX300 eye-trackers in standard laboratory conditions (Tobii, 2010; see Dalrymple et al., 2018 for data from 8–11 year-old children). This offset could result from WebGazer inaccuracy or because participants are looking somewhere else. We believe that the latter explanation does not play a substantial role, as: i) piloting the task over Zoom showed children directing their eyes towards target stimuli; ii) children similarly directed their eyes towards visual cues in the Experiment 1 WebGazer calibration sequence; and iii) if inattention were the primary driver of this gap, we would expect to see a more random distribution of looks in the Figure 16 density plots. Furthermore, this larger offset is consistent with prior WebGazer studies with adults; for example, Semmelmann and Weigelt (2018) and Slim and Hartsuiker (2022) reported offsets of 18% and 30% of screen-size (respectively) in online fixation tasks.

In particular, WebGazer appears to have difficulty discriminating looks along the vertical axis. This was evidenced by elevated looks to the image or quadrant vertically adjacent to the target and by greater vertical than horizontal offsets between gaze and target location. The results of Experiment 2 suggest that this difficulty may be greater for stimuli appearing on the top half of the screen. We observed a similar (but less pronounced) pattern in the video data in Experiment 1. Poor vertical resolution could reflect three constraints. First, most computer screens are rectangles with a landscape orientation, thus vertical distances between stimuli are generally smaller than horizontal ones. Second, webcams are typically placed above the screen but centered on the left–right axis. Consequently, a left look will be in the opposite direction relative to the webcam from a right look. In contrast, looks to both the upper and the lower half of the screen will be downward relative to the webcam. Finally, while it is easy to encourage participants to center themselves relative to their screen on the left–right axis (by sliding their computer or chair), vertical position is variable and more difficult to control. Most adults sit with their eyes above the screen, and thus the WebGazer algorithm was presumably trained on data of this kind. Children, who are shorter but live in a world of artifacts scaled to adults, typically sit with their heads nearer to the level of the screen. This may explain why the vertical spread is greater for children in the WebGazer data.

In our study, we identified two factors that influence WebGazer's performance: calibration score and participant age. Higher calibration scores are associated with data patterns suggesting better gaze tracking. In Experiment 2, the distance between estimated looks and the true stimulus location was reduced for participants with higher mean calibration scores (see also Slim & Hartsuiker, 2022). In both Experiment 1 and the adult pilot experiment, the size of the cohort effect was larger in trials with higher scores on the preceding calibration check (see Supplementary Materials). These results highlight the potential utility of calibration thresholds as a means to improve data quality, though the threshold of 50% often used in adult WebGazer experiments (e.g., Slim & Hartsuiker, 2022, Experiment 2; Vos et al., 2022) may be too high a bar for younger child participants (see Table 2).

Participant age also seems to influence WebGazer accuracy. WebGazer's spatial resolution appears higher for adult participants compared to child participants: in Experiment 2, the Euclidean distance offset between estimated gaze location and the true target location was smaller for adults, and in the quadrant analysis, the adult data yielded a higher proportion of target quadrant looks. Moreover, the age of the child participants influenced both calibration score and Euclidean distance offset: calibration scores were higher and distance offsets were smaller for older children.

Age-related differences could reflect factors specific to WebGazer. For example, older children may be in a more optimal position for WebGazer, because they are generally taller and thus may be positioned more like adults. In addition, older children tend to have larger faces than younger children, which could facilitate WebGazer's pupil detection and gaze estimation algorithms. Alternatively, age-related differences could reflect differences between participants that are independent of the technology used to estimate gaze. For example, older children may be less susceptible to distraction and more likely to sit still throughout the duration of the task.

Temporal resolution

In addition to observing differences in the eye-tracking methods' spatial resolutions, we also found that the timing of effects was slower than expected when WebGazer was used. In Experiment 1, in the absence of cohort competition, WebGazer detected reliable preferences for the target in clusters starting 800ms after target word onset for the two-picture display and 1200ms after target onset for the four-picture display. In contrast, in the annotated video data, this preference emerged in clusters starting 500ms after target word onset in the two-picture display and 700ms after target word onset in the four-picture display, similar to the timing in laboratory-based studies (Sekerina & Brooks, 2007; Simmons, 2017). We also found delays in the timing of the phonemic cohort effects (discussed below).

The apparent lag is not limited to studies with linguistic stimuli: we observed comparable WebGazer fixation delays in Experiment 2. In in-lab settings, saccade latencies in response to perceptual stimuli take approximately 200–250ms for adults (e.g., Matin et al., 1993; Rayner, Slowiaczek et al., 1983; Saslow, 1967; Theeuwes et al., 1998; Walker et al., 2000; White et al., 1962) and ten to twelve year-old children (Q. Yang et al., 2002). Q. Yang et al. (2002) observed mean latencies of approximately 300–350ms for children between the ages of four-and-a-half to twelve years. In contrast, in Experiment 2, looks settled on the target location approximately 500ms after onset (see also Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2022, for evidence of fixation delays with WebGazer).

We can imagine two possible explanations for this lag, which are not mutually exclusive. First, WebGazer could detect the same eye-movements as other eye-tracking measures but do so later due to time-consuming steps in the execution of the algorithm. Second, the lag could be a side-effect of WebGazer's poorer signal-to-noise ratio: effect sizes at the onset of an eye-movement pattern are typically smaller, making differences more difficult to detect. The data to date suggest that both factors play a role. On the one hand, streamlining WebGazer's algorithm to remove unnecessary computations improves its temporal resolution (X. Yang & Krajbich, 2021), suggesting processing limitations result in temporal delays. On the other hand, the variability that we observed in the WebGazer estimates well after stimulus onset (Figure 16) demonstrates that the spatial signal has substantial noise. Since even more streamlined versions of the WebGazer algorithm produce smaller effects than in-lab baselines (Vos et al., 2022), we expect that they would also fail to detect the earliest and weakest effects. Critically, we did not see comparable delays in the simultaneously-collected video data (Experiment 1), demonstrating that these delays are due to properties of the WebGazer algorithm and its execution and not to the less controlled nature of web-based settings.

Using webcam eye-tracking to detect fine-grained linguistic effects

Our findings suggest that WebGazer is not well suited for studying small or fleeting effects in children, particularly in the typical quadrant-based visual-world display. This was most clearly demonstrated by our analyses of the phonemic cohort effect in Experiment 1. In

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

the annotated webcam video data, we found significant cohort effects in both the two- and four-image displays, despite a sample of just 13 participants in each experiment. In contrast, even though our WebGazer sample contained more than twice as many participants (N=32 per experiment), WebGazer only detected evidence of a cohort effect in the two-image display. Moreover, the cluster window containing the effect was later and shorter than that in the video data (extending from 900–1099ms vs. 700–1099ms). Prior studies with adults have similarly observed WebGazer effects emerging later than in-lab baselines (Degen et al., 2021; Slim & Hartsuiker, 2022) as well as effects that are smaller and/or noisier than in-lab counterparts (Degen et al., 2021; Vos et al., 2022; Slim & Hartsuiker, 2022).

We conducted a series of power simulations using the {mixedpower} package v0.1.0 (Kumle et al., 2021) to assess the relative effect sizes in our webcam video and WebGazer data (see Supplementary Materials). We analyzed the likelihood of competitor image looks in the time windows where the Experiment 1 cluster analyses identified a difference between the two conditions (700–1099ms in Experiment 1A; 600–999ms in Experiment 1B).

In Experiment 1A (the two-image display), the effect of condition was larger in the webcam video data (β=-1.12, z=-3.52, p<0.001) than in the WebGazer data (β=-0.29, z=-1.83, p=0.07). In fact, the WebGazer effect was only 26% as large as the webcam video effect (as measured by the standardized beta coefficients). Given the sample sizes that we had, the observed power was 94% in the video data (N=13) and 45% in the WebGazer data (N=32). To achieve 94% power with the WebGazer data, the sample size would have to be increased to approximately 125 participants. To achieve power of at least 80% in the WebGazer data, the sample size would have to be increased to approximately 65 participants (power=80%). In contrast, reaching 80% power in the video data requires only seven participants (power=81%). In short, these simulations suggest that to achieve comparable power, a WebGazer study of this kind would require almost ten times as many participants as a study relying on webcam video annotation.

In Experiment B (the four-image display), the effect of condition was significant in the webcam video data (β=-1.08, z=-3.97, p<0.0001), with an observed power of 98% (N=13). To reach at least 80% power for an effect of this size required only five participants (power=82%). The effect of condition was not reliable in the WebGazer data (β=-0.03, z=-0.18, p=0.86; observed power 5% for N=32). If we assume that the true effect in the WebGazer data was 25% the size of the effect in video data, then a sample of approximately 120 participants would be required to achieve power greater than 80% (power=84%). This sample is 24 times the required minimum for the video data effect size. This conjecture is based on the relative effect sizes in Experiment 1A, though it is of course possible that the true effect size for WebGazer is considerably larger, or smaller, than our estimate.

In sum, our results suggest that webcam video annotation is a far more sensitive means of detecting the kind of fine-grained eye-movement effects that are relevant to many child language researchers. WebGazer estimation may be better suited to detecting fairly long-lasting effects in which the primary outcome measure is which part of the screen participants fixated on.

Recommendations for practice and directions for future research

Our results suggest that while both webcam video annotation and WebGazer estimation can be used with child participants in web-based tasks, the two methods have different advantages and disadvantages.

Webcam video annotation has better spatiotemporal accuracy than WebGazer (drastically reducing the amount of noise in our child data), making the method better suited to detecting the temporally-sensitive, fine-grained looking patterns assessed in studies of real-time language processing. Collecting webcam video data over Zoom requires relatively little technical expertise, as the experiment itself can be built and run in any software; the experiment can either be run on the participant's computer (as in Experiment 1) or displayed from the experimenter's computer using Zoom's screen sharing function (as in unpublished work by Anthony Yacovone, personal communication). It is possible to collect webcam video for gaze annotation in unsupervised web experiments using Zoom (Slim et al., 2022) or other webcam recording functions (e.g., via PCIbex; Ovans, 2022). However, the hand annotation process is time consuming (in Experiment 1, annotating a seven second video took approximately one minute), and the resulting gaze location estimates are relatively coarse-grained (representing regions of the display instead of coordinate estimates).

WebGazer's gaze coding, on the other hand, is automatic, reducing the data processing burden on the researcher. It can be used to obtain either gaze coordinate estimates or binary looks to relevant screen locations, and the data are saved in text format, thus helping to maintain participant privacy and requiring less storage space than video recordings. Our results suggest that it is possible to achieve similar target look resolution with WebGazer in quadrant-based analyses in both supervised and unsupervised web-based studies. In addition, WebGazer is free to use and has implementations in popular frameworks for web-based research. However, use of these implementations often requires working proficiency in programming languages, and implementations may not be compatible with all web-browsers. Furthermore, WebGazer's low spatiotemporal accuracy makes it more difficult to detect fine-grained effects with sufficient resolution and power. The sample sizes required to detect such effects with sufficient power are much larger than for webcam video annotation (10x the size or greater). These sample sizes may be prohibitively large for experiments targeting smaller effects. Nevertheless, WebGazer was able to detect looks towards targets in both of our experiments, suggesting that it is suitable for tasks that require spatial discrimination of robust looking patterns.

It is possible that the quality of data collected with WebGazer would be improved by having participants complete the experiment in the same environment or with the same computer (e.g., Özgoy et al., 2023; Semmelmann & Weigelt, 2018) – for instance, if a researcher uses a laptop as a mobile lab. However, recent work in our lab with Mieke Slim and Anthony Yacovone suggests that the limitations of WebGazer persist under more controlled conditions; in a comparison of an infrared eye-tracker, WebGazer, and webcam video annotation, we found no substantial differences in eye-movement effects when the two webcam methods were applied in the lab or in a web-based setting (where participants completed the experiment from their own computers).

Researchers should consider these trade-offs when deciding whether to conduct eye-tracking studies online and which gaze estimation method to use. For researchers interested in using WebGazer for online studies, we have several recommendations:

1) When designing the task, do not rely on vertical distinctions between critical stimuli. Consider simplifying the task to involve a two-image display or place critical stimuli on different halves of the screen in quadrant-based designs. Looks to diagonally-adjacent stimuli may be most easily discriminated (see Experiment 2).

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

2) When determining sample size, assume a 50–75% reduction in effect size relative to in-lab effects. Specifically, we found that the effects observed using WebGazer were roughly 25% as large (for the Experiment 1A cohort effect) to 45% as large (for horizontal target-side looks in the Experiment 1 control trials) as in the webcam video data, which produced effects of roughly the same magnitude as prior in-lab studies. The estimated reduction in effect size for WebGazer appears to vary based on effect type (short-lived, small effects vs. long-lasting fixations). Future work should investigate the performance of webcam eye-tracking methods in detecting various types of effects in order to provide more accurate recommendations for estimating expected effect sizes.
3) When planning the analysis, consider the likelihood of temporal delays in effect emergence. To account for such delays, researchers should shift or widen their planned analysis window appropriately or use an analysis method that does not assume a precise effect time window (e.g., cluster permutation analyses).
4) Consider setting calibration thresholds and/or including recalibration checkpoints to encourage participants to remain in an optimal position for WebGazer. Based on the data in Figure 14, we tentatively recommend a calibration threshold of at least 30% (though thresholds may need to be higher for smaller effects and/or more complicated displays).[12] To help improve WebGazer performance, ask parents to adjust their child's distance from the computer, the camera angle, and room lighting as necessary so that the participant's eyes can easily be seen in the webcam video feed at the onset of the calibration sequence.

For researchers interested in using webcam video annotation, we have the following recommendations:

1) Consider placing critical stimuli on different halves of the screen. While hand annotators were better at distinguishing quadrant looks than WebGazer in Experiment 1, horizontal differences are still easier for annotators to discriminate.
2) Re-center participant gaze with a central fixation prior to the onset of the experimental stimuli; having the gaze begin in the center of the screen makes it easier to identify in which direction looks are launched.
3) If using Zoom to record webcam video, utilize the gallery view layout for the recording (as opposed to the active speaker view) and hide non-video participants. This will ensure that the participant's webcam stream is present in the recording throughout the entire duration of the experiment. If the experimenter(s) turn off their video after starting the recording, the participant's face will be the only recorded view (note that at the time of writing, turning the experimenter video off prior to starting the recording causes Zoom to default to recording active speaker view). Zoom allows for simultaneous recordings in multiple layouts (e.g., screen recording, screen recording + thumbnail speaker view), which may be useful for aligning the recorded gaze data to trial onsets in the experiment.

[12] See Supplementary Materials for an analysis of the cohort effect in the Experiment 1 WebGazer data restricted to trials that meet this calibration threshold. In Experiment 1A, the cohort effect cluster increased in size (800–1199ms after target onset; z-sum=9.42, p<0.01) relative to our original analysis; there was still no cohort effect identified in the Experiment 1B WebGazer data. See Supplementary Materials for additional analyses relating trial calibration score to the size of the phonemic cohort effect in Experiments 1A and 1B.

4) If using teleconferencing software like Zoom to collect screen and/or webcam video recordings, test the available functions and settings for recordings. Some functions (e.g., Zoom's optimize for video function) may produce unexpected delays in the audio–visual sync within recordings.
5) Consider using a combination of visual and auditory prompts to identify trial onsets within the Zoom recordings; this will allow researchers to recover trial onsets should there be any issues with audio–visual synchrony in the recording.
6) Make sure that participant and/or experimenter has a way to view the participant webcam stream prior to the start of the experiment (e.g., through Zoom teleconference or by showing a video preview in PCIbex) so that the participant can adjust positioning and lighting to ensure that their eyes are visible in video recording.

While this work provides a starting point for evaluating online eye-tracking research with children, much remains to be done. For instance, future work should compare webcam-based eye-tracking methods to traditional high-end in-lab eye-trackers and should further assess the feasibility of running unsupervised web-based experiments with children. Despite the success of the Experiment 2 fixation task, we know very little about the limits of unsupervised tasks, particularly those with more complicated designs. In addition, while we observed age-related differences in WebGazer accuracy, it is unclear what is driving those differences, the extent to which they might influence the detection of linguistic effects, and whether we should expect similar differences in annotated webcam videos. Finally, as improvements continue to be made to automatic gaze-coding algorithms, their performance with child populations will need to be re-assessed.

Conclusion

We have demonstrated in two experiments that it is possible to run web-based visual-world studies with school-aged children in both supervised and unsupervised experimental settings. We tested two webcam eye-tracking methods and found that they are differentially suitable for detecting different kinds of effects. While both methods can discriminate looks to a target (albeit with different levels of accuracy), we found that WebGazer is not well-suited to detecting effects that require a high level of spatiotemporal accuracy (see Slim & Hartsuiker, 2022 for a similar conclusion). In contrast, frame-by-frame annotation of gaze direction from webcam videos provided sufficient spatial and temporal resolution to detect a fleeting and subtle effect typical of those studied by child language researchers. We anticipate that webcam eye-tracking will continue to improve as researchers develop tools, experimental protocols, and practices that are more precise, accurate, and efficient. We hope that these improvements will allow child language researchers to take advantage of the benefits of large-scale web-based experimentation for eye-tracking research.

Data availability statement. Data, analysis code, and Supplementary Materials are available from https://osf.io/hmeyb/.

Acknowledgements. Many thanks to Beatriz Leitão, Emily Liu, Mikaela Martin, Danielle Novak, Madeleine Presgrave, and Anthony Yacovone for their assistance with stimulus preparation, data collection, and/or data processing for this project. We are additionally grateful to Anthony Yacovone and Mieke Slim for sharing their thoughts on webcam eye-tracking methods, to Joshua Cetron and Patrick Mair for sharing their thoughts on the analyses, to the anonymous reviewers of this article for their helpful comments, and to all of the families and individuals who took the time to participate in this research. This work was supported by an internal grant from the Harvard University Psychology Department.

Competing interest. The authors declare none.

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. https://doi.org/10.1006/jmla.1997.2558

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

Bates, D., **Mächler, M.**, **Bolker, B.**, & **Walker, S.** (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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. https://doi.org/10.1016/j.jecp.2012.01.005

Brouwer, S., **Özkan, D.**, & **Küntay, A.** (2019). Verb-based prediction during language processing: The case of Dutch and Turkish. *Journal of Child Language*, **46**(1), 80–97. https://doi.org/10.1017/S0305000918000375

Contemori, C., **Carlson, M.**, & **Marnis, T.** (2018). On-line processing of English which-questions by children and adults: A visual world paradigm study. *Journal of Child Language*, **45**(2), 415–441. https://doi.org/10.1017/S0305000917000277

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*, **6**(1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-X

Cooper-Cunningham, R., **Charest, M.**, **Porretta, V.**, & **Järvikivi, J.** (2020). When Couches Have Eyes: The Effect of Visual Context on Children's Reference Processing. *Frontiers in Communication*, **5**, Article 576236. http://doi.org/10.3389/fcomm.2020.576236

Dahan, D., & **Gaskell, M. G.** (2007). The temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. *Journal of Memory and Language*, **57**(4), 483–501. https://doi.org/10.1016/j.jml.2007.01.001

Dahan, D., **Magnuson, J. S.**, & **Tanenhaus, M. K.** (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, **42**(4), 317–367. https://doi.org/10.1006/cogp.2001.0750

Dahan, D., & **Tanenhaus, M. K.** (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **30**(2), 498–513. https://doi.org/10.1037/0278-7393.30.2.498

Dalrymple, K. A., **Manner, M. D.**, **Harmelink, K. A.**, **Teska, E. P.**, & **Elison, J. T.** (2018). An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Frontiers in Psychology*, **9**, Article 803, https://doi.org/10.3389/fpsyg.2018.00803

Degen, J., **Kursat, L.**, & **Leigh, D.** (2021). Seeing is believing: Testing an explicit linking assumption for visual world eye-tracking in psycholinguistics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, **43**(1), 1500–1506. https://escholarship.org/uc/item/6182t9jb

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, **47**(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Desroches, A. S., **Joanisse, M. F.**, & **Robertson, E. K.** (2006). Specific phonological impairments in dyslexia revealed by eyetracking. *Cognition*, **100**(3), B32–B42. https://doi.org/10.1016/j.cognition.2005.09.001

Duñabeitia, J. A., **Crepaldi, D.**, **Meyer, A. S.**, **New, B.**, **Pliatsikas, C.**, **Smolka, E.**, & **Brysbaert, M.** (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, **71**(4), 808–816. https://doi.org/10.1080/17470218.2017.1310261

Erel, Y., **Shannon, K. A.**, **Chu, J.**, **Scott, K.**, **Kline Struhl, M.**, **Cao, P.**, **Tan, X.**, **Hart, P.**, **Raz, G.**, **Piccolo, S.**, **Mei, C.**, **Potter, C.**, **Jaffe-Dax, S.**, **Lew-Williams, C.**, **Tenenbaum, J.**, **Fairchild, K.**, **Barmano, A.**, & **Liu, S.** (2022, May 1). *iCatcher+: Robust and automated annotation of infant's and young children's gaze direction from videos collected in laboratory, field, and online studies.* PsyArXiv. https://doi.org/10.31234/osf.io/up97k

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

Farris-Trimble, A., & **McMurray, B.** (2013). Test-retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. *Journal of Speech, Language, and Hearing Research*, **56**(4), 1328–1345. https://doi.org/10.1044/1092-4388(2012/12-0145)

Fields, E. C., & **Kuperberg, G. R.** (2019). Having your cake and eating it too: Flexibility and power with mass univariate statistics for ERP data. *Psychophysiology*, **57**(2), Article e13468. https://doi.org/10.1111/psyp.13468

Fraser, A., **Gattas, S. U.**, **Hurman, K.**, **Robison, M.**, **Duta, M.**, & **Scerif, G.** (2021, June 22). *Automated gaze direction scoring from videos collected online through conventional webcam.* PsyArXiv. https://doi.org/10.31234/osf.io/4dmjk

Gaston, P., **Lau, E.**, & **Phillips, C.** (2020, December 4). *How does(n't) syntactic context guide auditory word recognition?* PsyArXiv. https://doi.org/10.31234/osf.io/sbxpn

Gelman, A., & **Hill, J.** (2007). *Data analysis using regression and multilevel hierarchical models* (Vol. **1**). Cambridge University Press.

Griffin, Z., & **Bock, K.** (2000). What the eyes say about speaking. *Psychological Science*, **11**(4), 274—279. https://doi.org/10.1111/1467-9280.00255

Groppe, D. M., **Urbach, T. P.**, & **Kutas, M.** (2011). Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. *Psychophysiology*, **48**(2), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

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. https://doi.org/10.1002/aur.1487

Henrich, J., **Heine, S. J.**, & **Norenzayan, A.** (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, **33**(2-3), 61–83. https://doi.org/10.1017/s0140525x0999152x

Huang, Y. T., & **Snedeker, J.** (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, **45**(6), 1723–1739. https://doi.org/10.1037/a0016704

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. https://doi.org/10.1016/j.jml.2007.02.001

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

Ito, A., **Pickering, M. J.**, & **Corley, M.** (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, **98**, 1–11. https://doi.org/10.1016/j.jml.2017.09.002

Kampa, A., & **Papafragou, A.** (2020). Four-year-olds incorporate speaker knowledge into pragmatic inferences. *Developmental Science*, **23**(3), Article e12920. https://doi.org/10.1111/desc.12920

Kumle, L., **Võ, M. L.-H.**, & **Draschkow, D.** (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavioral Research Methods*, **53**(6), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Li, X., **Li, X.**, & **Qu, Q.** (2022). Predicting phonology in language comprehension: Evidence from the visual world eye-tracking task in Mandarin Chinese. *Journal of Experimental Psychology: Human Perception and Performance*, **48**(5), 531–547. https://doi.org/10.1037/xhp0000999

Magnuson, J. S., **Tanenhaus, M. K.**, **Aslin, R. N.**, & **Dahan, D.** (1999). Spoken word recognition in the visual world paradigm reflects the structure of the entire lexicon. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, 331–336.

Matin, E., **Shao, K. C.**, & **Boff, K. R.** (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, **53**(4), 372–380. https://doi.org/10.3758/BF03206780

McMurray, B., **Danelz, A.**, **Rigler, H.**, & **Seedorff, M.** (2018). Speech categorization develops slowly through adolescence. *Developmental Psychology*, **54**(8), 1472–1491. https://doi.org/10.1037/dev0000542

Meyer, A. S., **Sleiderink, A. M.**, & **Levelt, W. J. M.** (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, **66**(2), B25—B33. https://doi.org/10.1016/s0010-0277(98)00009-2

Ovans, Z. (2022). *Developmental parsing and cognitive control* (Doctoral dissertation). Retrieved from https://doi.org/10.13016/en2r-ce6z

Özge, D., **Kornfilt, J.**, **Maquate, K.**, **Küntay, A. C.**, & **Snedeker, J.** (2022). German-speaking children use sentence-initial case marking for predictive language processing at age four. *Cognition*, **221**, Article 104988. https://doi.org/10.1016/j.cognition.2021.104988

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)*, **25**(1), 3839–3845.

Paul, P., **Ziegler, J.**, **Chalmers, E.**, & **Snedeker, J.** (2019). Children and adults successfully comprehend subject-*only* sentences online. *PLoS ONE*, **14**(1), Article e0209670. https://doi.org/10.1371/journal.pone.0209670

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

Rayner, K., **Slowiaczek, M. L.**, **Clifton, C.**, & **Bertera, J. H.** (1983). Latency of sequential eye movements: Implications for reading. *Journal of Experimental Psychology: Human Perception and Performance*, **9**(6), 912–922. https://doi.org/10.1037/0096-1523.9.6.912

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

Rigler, H., **Farris-Trimble, A.**, **Greiner, L.**, **Walker, J.**, **Tomblin, J. B.**, & **McMurray, B.** (2015). The slow developmental time course of real-time spoken word recognition. *Developmental Psychology*, **51**(12), 1690–1703. https://doi.org/10.1037/dev0000044

Rossion, B., & **Pourtois, G.** (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, **33**(2), 217– 236. https://psycnet.apa.org/doi/10.1068/p5117

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

Sassenhagen, J., & **Draschkow, D.** (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, **56**(6), Article e13335. https://doi.org/10.1111/psyp.13335

Sekerina, I. A., & **Brooks, P. J.** (2007). Eye movements during spoken word recognition in Russian children. *Journal of Experimental Child Psychology*, **98**(1), 20–45. https://doi.org/10.1016/j.jecp.2007.04.005

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

Simmons, E. S. (2017). *The timecourse of phonological competition in spoken word recognition: A comparison of adults and very young children* (Master's thesis). Retrieved from https://opencommons.uconn.edu/gs_theses/1156/.

Slim, M. S., & **Hartsuiker, R. J.** (2022). Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js. *Behavioral Research Methods.* https://doi.org/10.3758/s13428-022-01989-z

Slim, M. S., **Hartsuiker, R. J.**, & **Snedeker, J.** (2022). *The real-time resolution of quantifier scope ambiguity.* Paper presented at the 22nd ESCOP Conference, Université de Lille, Lille, France.

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. https://doi.org/10.1016/j.cogpsych.2004.03.001

SR Research. (2021). *EyeLink® 1000 Plus Brochure.*

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. https://doi.org/10.1126/science.7777863

Theeuwes, J., **Kramer, A. F.**, **Hahn, S.**, & **Irwin, D. E.** (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science*, **9**(5), 379–385. https://doi.org/10.1111/1467-9280.00071

Tobii (2010). *Tobii TX300 Eye Tracker.*

Tobii (2021). *Pro Spectrum User Manual.*

Trueswell, J. C., **Sekerina, I.**, **Hill, N. M.**, & **Logrip, M. L.** (1999). The kindergarten-path effect studying on-line sentence processing in young children. *Cognition*, **73**(2), 89–134. https://doi.org/10.1016/s0010-0277(99)00032-3

https://doi.org/10.1017/S0305000924000175 Published online by Cambridge University Press

Valenti, R., **Staiano, J.**, **Sebe, N.**, & **Gevers, T.** (2009). Webcam-based visual gaze estimation. *International Conference on Image Analysis and Processing – ICIAP 2009*, 5716, 662–671. https://doi.org/10.1007/978-3-642-04146-4_71

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

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

Walker, R., **Walker, D. G.**, **Husain, M.**, & **Kennard, C.** (2000). Control of voluntary and reflexive saccades. *Experimental Brain Research*, **130**(4), 540–544. https://doi.org/10.1007/s002219900285

Weighall, A. R., **Henderson, L. M.**, **Barr, D. J.**, **Cairney, S. A.**, & **Gaskell, M. G.** (2017). Eye-tracking the time-course of novel word learning and lexical competition in adults and children. *Brain and Language*, **167**, 13–27. https://doi.org/10.1016/j.bandl.2016.07.010

White, C. T., **Eason, R. G.**, & **Bartlett, N. R.** (1962). Latency and duration of eye movements in the horizontal plane. *Journal of the Optical Society of America*, **52**(2), 210–213. https://doi.org/10.1364/josa.52.000210

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

Yacovone, A., **Shafto, C. L.**, **Worek, A.**, & **Snedeker, J.** (2021). Word vs. world knowledge: A developmental shift from bottom-up lexical cues to top-down plausibility. *Cognitive Psychology*, **131**, Article 101442. https://doi.org/10.1016/j.cogpsych.2021.101442

Yang, Q., **Bucci, M. P.**, & **Kapoula, Z.** (2002). The latency of saccades, vergence, and combined eye movements in children and in adults. *Investigative Ophthalmology & Visual Science*, **43**(9), 2939–2949.

Yang, X., & **Krajbich, I.** (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision Making*, **16**(6), 1485–1505. https://doi.org/10.1017/S1930297500008512

Zehr, J., & **Schwarz, F.** (2018). PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832

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

Cite this article: Kandel, M., & Snedeker, J. (2024). Assessing two methods of webcam-based eye-tracking for child language research. *Journal of Child Language* 1–34, https://doi.org/10.1017/S0305000924000175