Online visual world eye-tracking using webcams[1]

Mieke Sarah Slim and Robert J. Hartsuiker; Ghent University, Belgium
mieke.slim@ugent.be

Web-based experimentation has become a viable and efficient alternative to lab-based experimentation in psycholinguistics (e.g., Gibson, Piantadosi, & Fedorenko, 2011). However, some experimental paradigms, such as eye-tracking, require stationary equipment and are therefore bound to the lab. With the arrival of WebGazer, a webcam-based eye-tracking algorithm, web-based eye-tracking has become possible. WebGazer consists of a pupil detector, which locates the eyes based on the webcam stream, and a gaze estimator, which estimates the gaze location of the participants using regression analysis (Papoutsaki et al., 2016). This algorithm opens up the possibility for online eye-tracking studies in psycholinguistic research. In this study, we investigated whether web-based eye-tracking is a viable option to conduct visual world studies. We tested this question in two experiments: A fixation task (Experiment 1) and a replication of a visual world study (Experiment 2). Both experiments were implemented using the PCIbex library (Zehr & Schwarz, 2018), which contains an eye-tracker element that uses the WebGazer algorithm.

The main goal of the fixation task in Experiment 1 was to gain insight in temporal and spatial accuracy of the webcam eye-tracker without testing any linguistic mechanisms that impact eye fixations (cf. Semmelmann & Weigelt, 2018). Moreover, we tested to what extent the calibration precision of the eye-tracker impacts the accuracy of the results. In this task, the participants (*n* = 50, recruited via Prolific) looked at a fixation cross that appeared in one of thirteen positions on the screen for 1500 ms (Figure 1). Estimated fixations landed on the stimulus cross after roughly 500 ms (Figure 2A). Since it typically takes approximately 200 ms to launch a saccade (e.g., Matin, Shao & Boff, 1993), there seems to be a systematic delay of about 300 ms in the eye-tracking recordings. Focusing on spatial accuracy of the data, the distance between the middle of fixation cross and the estimated fixation location was roughly 30% of the screen size (Figure 2A). However, spatial accuracy was modulated by calibration precision: The spatial accuracy improved if the eye-tracker was better calibrated (Figure 2B).

In Experiment 2, we test the viability of web-based eye-tracking using the visual world paradigm by replicating a lab-based visual world experiment by Dijkgraaf, Hartsuiker, and Duyck (2017) in an online-setting. This experiment tested lexical-semantic predictive processing at the verb in language comprehension, an effect that is often found in visual world studies (e.g., Altmann & Kamide, 1999; Kamide, Altmann & Haywood, 2003). The participants listened to simple transitive sentences like *Mary stole a letter* while they looked at four images that were displayed in the four quadrants of the screen (Figure 3). One of the four objects on the display depicted the object phrase of the auditory stimulus (the target image). These stimuli were presented in two experimental conditions: the *neutral* condition and the *constrained* condition. In the neutral condition, all four objects in the display were appropriate after the verb. In the constrained condition, however, only the target image was appropriate after the verb. Dijkgraaf et al.'s data showed that the participants (30 native speakers of English) tend to look at the target image prior to the onset of the object noun in the constrained condition. In the neutral condition, the participants fixated on the target image after the object noun onset. In Experiment 2 (*n* = 90, recruited via Prolific), we replicated this effect in a web-based setting. However, cluster permutation analyses on these data did reveal a delay in the effect of roughly 300 ms relative to Dijkgraaf et al.'s results (Figure 4).

Altogether, the results of Experiment 1 and Experiment 2 show that web-based eye-tracking is promising, but the spatial and temporal resolution of online eye-tracking is considerably poorer compared to in-lab testing using an eye-tracking device. Therefore, online eye-tracking may not be suitable for paradigms that require a close spatial and temporal resolution (e.g., eye-tracking while reading). Nevertheless, our results show that online eye-tracking is accurate enough to detect effects of lexical-semantic constraints using the visual world paradigm.

[1] This study is pre-registered: https://osf.io/yfxmw/registrations

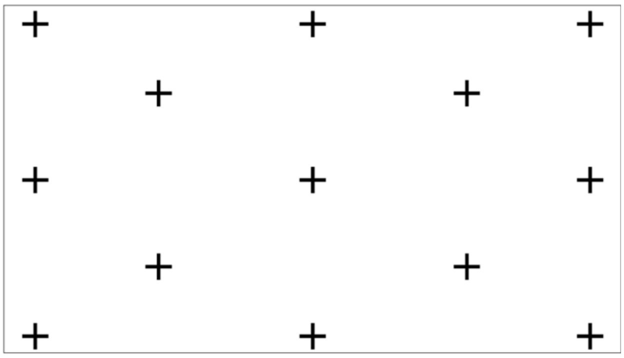

Figure 1. The thirteen possible positions of the stimuli on the computer screen in Experiment 1.

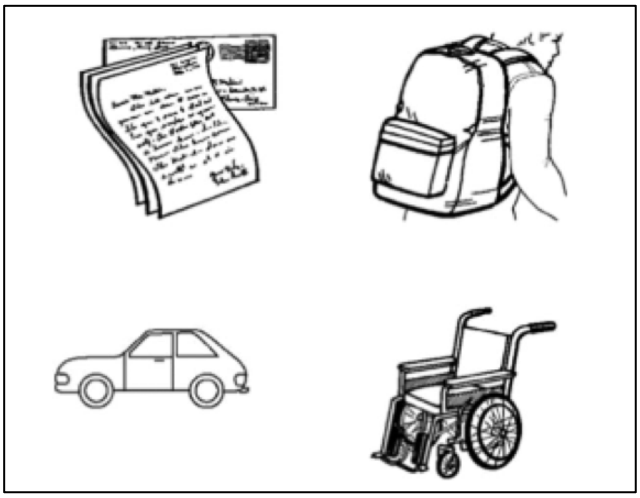

Figure 3. Example of a visual scene used in Dijkgraaf, Hartsuiker, and Duyck (2017) and in Experiment 2. This display is used in both the constrained condition (in which case the sentence *Mary reads a letter* was played) and in the neutral condition (in which case the sentence *Mary steals a letter* was played).

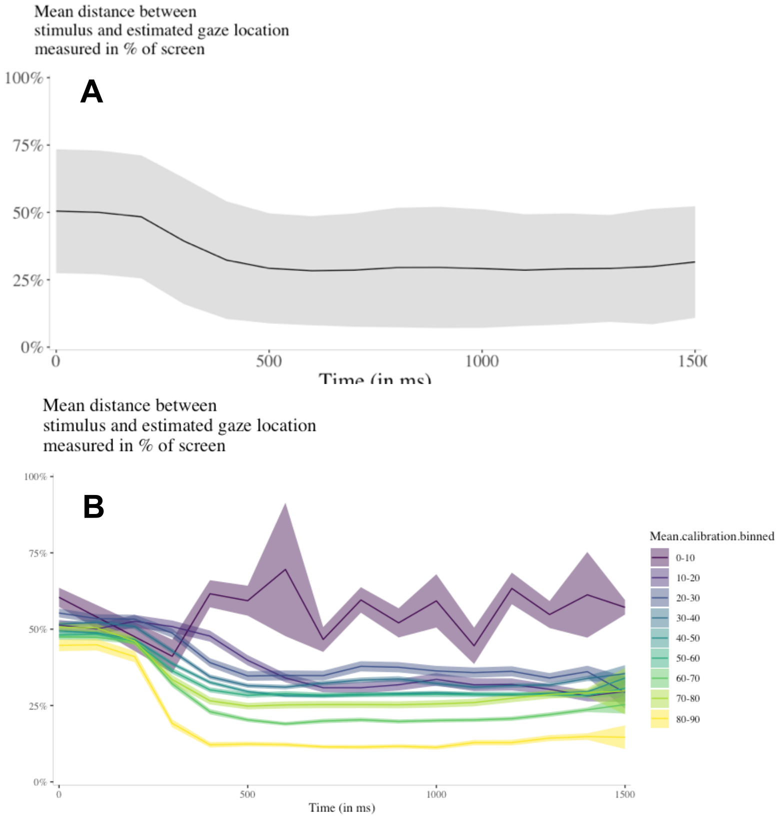

Figure 2. A: The mean distance between estimated gaze location and the middle of the stimulus as a function of time. The standard deviation is represented in grey. Note that after 500 ms, the participants more-or-less settle their gazes on the stimulus. B: The mean distance as a function of time, divided over calibration precision. Calibration precision was calculated on a scale from 0 (very badly calibrated) to 100 (perfectly calibrated). In this plot, calibration precision (per participant) is binned in 10-point bins. Importantly, this figure suggests that a more precise calibration improves the spatial resolution of the data.

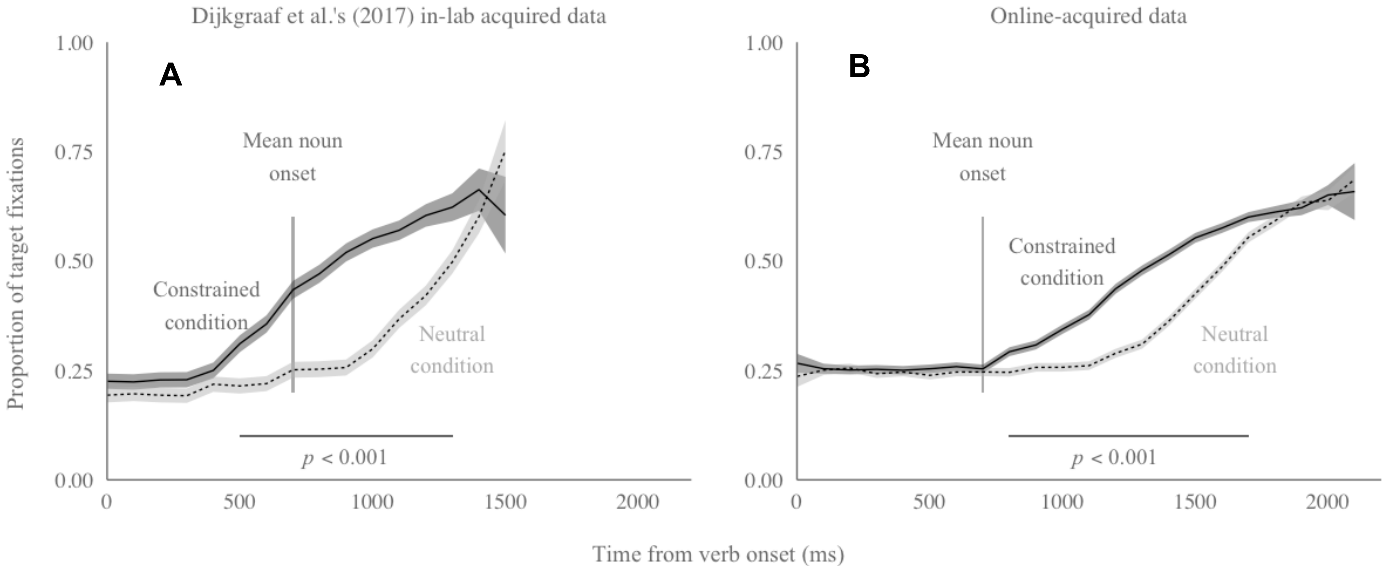

Figure 4: Results from Dijkgraaf et al.'s (2017) original experiment (A) and preliminary results of Experiment 2 (B). Cluster permutation analyses revealed that the difference between the neutral and the constrained condition was reliable between 500 and 1400 ms, as indicated by the shaded area ($p < 0.001$) in Dijkgraaf et al.'s results. This effect was reliable between 800 and 1700 ms in the results of Experiment 2 ($p < 0.001$). This suggests that there is a delay of 400 ms in the results of Experiment 2.

NB: Note that the eye-tracking recordings were longer in the present study than in Dijkgraaf et al,'s experiment, because we expected possible delayed latencies based on the results of Experiment 1.

References: Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73(3), 247-264. Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. Bilingualism: Language and Cognition, 20(5), 917-930. Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. Language and Linguistics Compass, 5(8), 509-524. Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer : Scalable webcam eye tracking using user interactions. International Joint Conference on Artificial Intelligence. Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and language, 49(1), 133-156. Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. Perception & Psychophysics, 53, 372–380. Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. Behavior Research Methods, 50(2), 451-465. Zehr, J., & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832