Eye-tracking on the web: lessons learned from replicating 6 experiments

First Author[1] & Ernst-August Doelle[1,2]

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. First Author: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ernst-August Doelle: Writing - Review & Editing.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: my@email.com

Abstract

ADD LATER

Keywords: keywords

Word count: X

Eye-tracking on the web: lessons learned from replicating 6 experiments

Intro stuff:

- Eye-tracking as a key method in cognitive science research
- Online data collection is more and more popular & let's us ask new questions
- But, concerns over quality + little known about eye-tracking online

Present work

In the present work, we attempted to replicate six eye-tracking studies from the cognitive science literature using the `jsPsych` platform and `webgazer.js` plug-in. The goal was to examine the strengths and weaknesses of webcam eye-tracking for common paradigms in cognitive science. The studies were chosen to cover a variety of topic areas (e.g., memory, decision-making, psycholinguistics) and paradigms (two halves of the screen, visual world paradigm with four quadrants, visual world paradigm with "natural" scenes). . . .

Experiment 1

The first study was a replication attempt of Altmann and Kamide (1999).

Methods

All stimuli, experiment scrips, data, analysis scripts, and pre-registration are available on the Open Science Framework at https://osf.io/s82kz. All participants provided informed consent and this study was approved by the Vassar College Institutional Review Board.

Participants. Participants for this experiment were sampled from a wide pool of Prolific users who are fluent in English and were paid for their participation. Our sample size of participants was determined by the total run time of our experiment, ~10 minutes, and the allotted funding that was endowed to us by the Vassar College Cognitive Science Department. From this information, we calculated a reasonable number of participants we could afford to compensate on Prolific, and we ended up with a sample size of 60 participants. For unknown reasons, 2 of the subjects' results were not recorded, so in the analysis, we worked with data collected from 58 participants.

Procedure. Participants completed the experiment remotely and entirely online on the platform Prolific. During the experiment, the participants viewed a screen and were simultaneously presented with a visual image and a corresponding audio recording of a spoken sentence. The visual stimuli were created through Canva and depict a subject accompanied by 4 to 5 objects in the scene. 16 of the devised stimuli were critical to our trial, and each of these images have 2 sentences associated with it. One of these sentences is in the restrictive condition, where the verb only applies to one object in the scene, and the other is in the nonrestrictive condition, where the verb could apply to all of the objects in the scene. To illustrate an example, reference Figure 1. This scene depicts a boy alongside a cake, ball, car, and train set. In the case of this particular image, the cake is the target object, so the two corresponding sentences to this image are, "The boy will eat the cake" (restrictive) and "The boy will move the cake" (nonrestrictive).

Materials. Therefore, for the 16 critical images, there are 16 control sentences where the verb does not constrain the target object and 16 restrictive sentences where the verb does constrain the target object. Each participant randomly received one sentence, restrictive or nonrestrictive, per scene. Of the 16 critical trials, each participant got 8 sentences that were restrictive and 8 that were nonrestrictive and the order of these were randomized. Trials were also designed so that participants had to input a keyboard response indicating "yes" or "no" as to whether the sentence relayed was feasible given the visual image. There were two practice trials to ensure that participants had a sound understanding of the instructions before they undertook the main portion of the experiment. In addition to the 16 critical images, we also devised an additional 16 filler images that are not pertinent to our data collection and analysis. The critical trials were also presented in a randomized order along with these 16 filler trials. Unlike the critical images, the filler images were accompanied by only one sentence that is unfeasible given it's corresponding scene. This was so that when participants were asked whether or not the scene was possible, the filler trials would always elicit the answer, "no." Despite recording participants' reaction time and keyboard response after each trial, we were specifically measuring for the participant's first fixation to the target object and distractors relative to the onset of the verb, the offset of the verb, onset of the post-verbal determiner, and onset of the target noun.

Eye-Tracking Calibration and Validation. Before initiating the experiment, participants were prompted to complete an eye-tracking calibration and validation procedure to ensure the data collected via the webcam-based eye tracking method was as accurate as possible. In order to allow the software to track where participants are looking, subjects were presented with a series of dots that appeared on the screen. Participants were then instructed to look at each dot as they appeared and click on it. By visually fixating and then clicking on one dot, it would then disappear and a new one would reappear in a different location on the screen. The calibration dots appeared in the central area of the screen where the visual stimuli would appear in order to ensure Web Gazer would be able to track eye movements to the relevant regions of interest. After completing this calibration, participants were then asked to go through the same steps of the calibration, except this time, they would have to just look at, not click, the dots as they appear on the screen in order to measure the accuracy of the calibration. This process completes the Web Gazer calibration and validation process.

Data pre-processing and analysis. We used R [Version 4.1.0; R Core Team (2021)] and the R-package *papaja* [Version 0.1.0.9997; Aust and Barth (2020)] for all our analyses.

Results

Replication.

here we will describe the analyses that are as close as possible to the original paper

Calibration.

here we will describe the analyses that correlate calibration quality with effect size at the individual

Discussion

Experiment 2

The second study was a replication attempt of Johansson and Johansson (2014).

Methods

Participants. Participants for this study were recruiting using the website Prolific. Specifically, participants had to be older than 18 and fluent in English, but aside from that, there was no restriction on demographic; participation was also anonymous. The only technology-based restriction was that each participant had to have a working webcam; we were able to accommodate for different screen sizes during data analysis. We analyzed the data of 59 participants, a number that was limited by budget constraints, but still 2.5x larger than the original sample size of 24, as suggested by Simonsohn (2015). We ended up excluding the data of 1 participant whose eye tracking data seemed to be blank, rendering us unable to analyze it.

Material.

Procedure. After a participant began the study, they would encounter an alert that the study would use their webcam to track eye movements, then initiate Webgazer, a program that logs predicted eye positions. There were two possible conditions they could begin with—the free-viewing condition and the fixed-viewing condition—which they were randomly assigned to. For both conditions, participants began with the encoding phase before moving to the recall phase. After going through both conditions, participants were asked a few survey questions and the experiment ended. For the encoding phase, participants were asked to remember the contents of the four quadrants of a grid, each with six distinct items themed around a certain category (we used humanoids, household objects, animals, and methods of transportation, as inspired by the example in Johansson & Johansson's original experiment). To do so, each of the four quadrants was presented to the participant one at a time. First, a list of the items in the quadrant were shown, then the items in the actual quadrant were shown. For each item in a quadrant, an audio file would play, asking the participant to use their arrow keys to identify which direction each item was facing (every item was facing a distinct direction to allow for statements like "The chair is facing left" to be viable). After the participant identified the direction of each item, they would have an additional 30 seconds to study and remember the name and orientation of each item in the quadrant. Then, after repeating this for each quadrant, the participant was shown the full grid of 24 items (six per grid) and had sixty seconds to further encode the name and orientation of each item. For the recall phase, participants had to respond to statements presented via audio files. Each statement was a true or false statement that fell into either an interobject or intraobject condition. Interobject statements were those that compared two different items in the grid (e.g. "The skeleton is to the left of the robot"), while intraobject statements were those that asked about the orientation of a single item (e.g. "The bus is facing right"). There were 48 total statements, with 24 interobject and 24 intraobject statements split evenly among the four quadrants. Participants were able to respond by pressing the 'F' key for false statements and 'T' for true ones. The difference between the free-viewing and fixed-viewing conditions was in what a participant could see on screen during the recall phase. During fixed-viewing, participants could only see a small cross in the center of the screen, and were asked at the start of the recall phase to focus their vision on just the cross. During free-viewing, participants saw an empty screen and were allowed to look wherever they wanted, without any particular instruction telling them where to look. In both cases, the mouse was obscured from the screen, making it so that the only visual stimulus for either condition was the cross in the fixed-viewing one. After a participant finished the first condition, they then moved on to repeating the experiment under the second condition, this time with a new grid of images. After completing this second condition, the participant was finished, and was asked to answer a few survey questions (such as whether they wore glasses or encountered any distractions). The methodology of this replication differed from Johansson & Johansson's original study in two key ways. First, the original study included two more conditions that were omitted from this replication for efficiency concerns. Those two conditions involved prompting the participant to look to an area on screen that matched one of the original quadrants while they responded to each statement, with the key here being that sometimes the prompt was in the same quadrant as the information being recalled, while sometimes it wasn't. We ended up not replicating this aspect of the study because it felt like too big of a task to program and playtest in the amount of time we had. The second major way this study differed from the original was that the original was conducted in- person and used an iView Red500 eye tracker, which was able to track eye movements incredibly closely. Due to obvious pandemic-related reasons, we conducted our replication online through Prolific, while using Webgazer, a software that tracks eye movements based on a participant's webcam. There were a few other minor ways in which our study differed (we used different grids of objects and different instructions), but we tried to remain as faithful as possible in terms of variables like timing, the physical appearance of the grids, and the style of the images.

Data analysis.

Results

Replication.

here we will describe the analyses that are as close as possible to the original paper

Calibration.

here we will describe the analyses that correlate calibration quality with effect size at the individual

Discussion

Experiment 3

The third study was a replication attempt of Manns, Stark, and Squire (2000).

Methods

Participants.

Material.

Procedure.

Data analysis.

Results

Discussion

Experiment 4

The fourth study was a replication attempt of Ryskin, Qi, Duff, and Brown-Schmidt (2017).

Methods

Participants.

Material.

Procedure.

Data analysis.

Results

Replication.

here we will describe the analyses that are as close as possible to the original paper

Calibration.

here we will describe the analyses that correlate calibration quality with effect size at the individual

Effects of ROIs.

here we will describe how results change depending on the size of the ROIs (using the image vs the screen quadrant)

Discussion

Experiment 5

The fifth study was a replication attempt of @??.

Methods

Participants.

Material.

Procedure.

Data analysis.

Results

Discussion

Experiment 6

The sixth study was a replication attempt of @??.

Methods

Participants.

Material.

Procedure.

Data analysis.

Results

Discussion

Combined Analyses

General Discussion

References

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

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Johansson, R., & Johansson, M. (2014). Look Here, Eye Movements Play a Functional Role in Memory Retrieval. *Psychological Science*, *25*(1), 236–242. https://doi.org/10.1177/0956797613498260

Manns, J. R., Stark, C. E. L., & Squire, L. R. (2000). The visual paired-comparison task as a measure of declarative memory. *Proceedings of the National Academy of Sciences*, *97*(22), 12375–12379. https://doi.org/10.1073/pnas.220398097

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

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