What paradigms can webcam eye-tracking be used for? Attempted replications of five "classic" cognitive science experiments

Joshua R. de Leeuw[1], Rachel Ryskin[2], Ariel N. James[3], Joshua K. Hartshorne[4], Haylee Backs[1], Nandeeta Bala[1], Laila Barcenas-Meade[1], Samata Bhattarai[1], Tessa Charles[1], Gerasimos Copoulos[1], Claire Coss[1], Alexander Eisert[1], Elena Furuhashi[1], Keara Ginell[1], Anna Guttman-McCabe[1], Emma (Chaz) Harrison[1], Laura Hoban[1], William A. Hwang[1], Claire Iannetta[1], Kristen M. Koenig[1], Chauncey Lo[1], Victoria Palone[1], Gina Pepitone[1], Margaret Ritzau[1], Yi Hua Sung[1], & Lauren Thompson[1]

[1] Cognitive Science Department, Vassar College

[2] Department of Cognitive & Information Science, University of California, Merced

[3] Psychology Department, Macalester College

[4] Department of Psychology & Neuroscience, Boston College

Author Note

The authors made the following contributions. Joshua R. de Leeuw: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing; Rachel Ryskin: Conceptualization, Formal analysis, Visualization, Writing - original draft, Writing - review & editing; Ariel N. James: Conceptualization, Formal analysis, Visualization, Writing - original draft, Writing - review & editing; Joshua K. Hartshorne: Conceptualization, Formal analysis, Visualization, Writing - original draft, Writing - review & editing; Haylee Backs: Investigation, Methodology, Software; Nandeeta Bala: Investigation, Methodology, Software; Laila Barcenas-Meade: Investigation, Methodology, Software; Samata Bhattarai: Investigation, Methodology, Software; Tessa Charles: Investigation, Methodology, Software; Gerasimos Copoulos: Investigation, Methodology, Software; Claire Coss: Investigation, Methodology, Software; Alexander Eisert: Investigation, Methodology, Software; Elena Furuhashi: Investigation, Methodology, Software; Keara Ginell: Investigation, Methodology, Software; Anna Guttman-McCabe: Investigation, Methodology, Software; Emma (Chaz) Harrison: Investigation, Methodology, Software; Laura Hoban: Investigation, Methodology, Software; William A. Hwang: Investigation, Methodology, Software; Claire Iannetta: Investigation, Methodology, Software; Kristen M. Koenig: Investigation, Methodology, Software; Chauncey Lo: Investigation, Methodology, Software; Victoria Palone: Investigation, Methodology, Software; Gina Pepitone: Investigation, Methodology, Software; Margaret Ritzau: Investigation, Methodology, Software; Yi Hua Sung: Investigation, Methodology, Software; Lauren Thompson: Investigation, Methodology, Software.

Correspondence concerning this article should be addressed to Joshua R. de Leeuw, 124 Raymond Ave, Poughkeepsie, NY 12604, USA. E-mail: jdeleeuw@vassar.edu

Abstract

Web-based data collection allows researchers to recruit large and diverse samples with fewer resources than lab-based studies require. Recent innovations have expanded the set of methodolgies that are possible online, but ongoing work is needed to test the suitability of web-based tools for various research paradigms. Here, we focus on webcam-based eye-tracking; we tested whether the results of five different eye-tracking experiments in the cognitive psychology literature would replicate in a webcam-based format. Specifically, we carried out five experiments by integrating two javascript-based tools: js.psych and a modified version of Webgazer.js. In order to represent a wide range of applications of eye-tracking to cognitive psychology, we chose two psycholinguistic experiments, two memory experiments, and a decision-making experiment. These studies also varied in the type of eye-tracking display, including screens split into halves (Exps. 3 and 5) or quadrants (Exps. 2 and 4), or composed scenes with regions of interest that varied in size (Exp. 1). Outcomes were mixed. The least successful replication attempt was Exp. 1; we did not obtain a condition effect in our remote sample (1a), nor in an in-lab follow-up (1b). However, the other four experiments were more successful, replicating a blank-screen effect (Exp. 2), a novelty preference (Exp. 3), a verb bias effect (Exp. 4), and a gaze-bias effect in decision-making (Exp. 5). These results suggest that webcam-based eye tracking can be used to detect a variety of cognitive phenomena, including those with sensitive time, although paradigms that require high spatial resolution should be adapted to coarser quadrant or split-half displays.

Keywords: eye-tracking, online, webcam, jsPsych, cognitive science

Word count: X

What paradigms can webcam eye-tracking be used for? Attempted replications of five "classic" cognitive science experiments

The use of eye-tracking to study cognition took off when Alfred Yarbus used suction cups to affix a mirror system to the sclera of the eye in order to monitor eye position during the perception of images (Yarbus, 1967). In one study, participants viewed a painting depicting multiple people in a complex interaction inside of a 19th century Russian home. Yarbus showed, among other things, that the scan paths and locations of fixations were largely dependent on the instructions given to participants (e.g., View the picture freely vs. Remember the position of the people and objects in the room). In other words, the cognitive processing that the individual is engaged in drives the visuo-motor system. Since these findings, eye-tracking has become a central method in cognitive science research Rayner (1998). For example, gaze location during natural scene perception is used to test theories of visual attention (e.g., Henderson & Hayes, 2017), and eye-movements during auditory language comprehension, using the "visual world paradigm," demonstrated the context-dependent and incremental nature of language processing (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

An important limitation of the eye-tracking methodology is that it has typically required costly equipment (eye-trackers can range in price from a few thousand dollars to tens of thousands of dollars), particular laboratory conditions (a quiet room with consistent indoor lighting conditions), and a substantial time investment (e.g., bringing participants into a laboratory one at a time). This limits who can conduct eye-tracking research – not all researchers have the necessary resources – and who can participate in eye-tracking research. Most eye-tracking study participants are from western, educated, industrialized, rich, and democratic [WEIRD; Henrich, Heine, and Norenzayan (2010)] convenience samples (but see Ryskin, Salinas, Piantadosi, & Gibson, 2023), which diminishes the generalizability of the findings and the scope of conclusions that can be drawn about human cognition. Likewise, the sample sizes for in-lab experiments are usually orders of magnitude smaller than what statisticians recommend (Nosek et al., 2022).

A robust solution to all these problems is online experiments, particularly with volunteer citizen scientists as participants (Gosling, Sandy, John, & Potter, 2010; Hartshorne, Leeuw, Goodman, Jennings, & O'Donnell, 2019; Li, Germine, Mehr, Srinivasan, & Hartshorne, 2024; Reinecke & Gajos, 2015). Historically, this option has not been available for eye-tracking. This began to shift with the widespread incorporation of cameras into computers; researchers have long used frame-by-frame analysis of video for low-resolution eyetracking (e.g., Snedeker & Trueswell, 2003). Unfortunately, this can be extremely time-intensive, making large-sample studies unrealistic. In recent years, image analysis has improved to the point where this work can be automated with reasonable accuracy (Burton, Albert, & Flynn, 2014; Papoutsaki et al., 2016; Skovsgaard, Agustin, Johansen, Hansen, & Tall, 2011; Zheng & Usagawa, 2018). However, webcam-based eyetracking only started to be used regularly in research with the advent of `Webgazer.js` (Papoutsaki et al., 2016), a webcam-based Javascript plug-in that works in the browser and which can be integrated with any Javascript web interface, including `jsPsych` (de Leeuw, 2015), `Gorilla` (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020), or `lab.js` (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2021).

Given the potential game-changing nature of webcam-based eye-tracking, a number of research groups have investigated how well it works. There are two potential limitations to webcam-based eye-tracking. First, the spatial and temporal resolution is less than what is achievable with an infrared system. Second, testing subjects over the Internet involves less control: subjects may be unable or unwilling to calibrate equipment, adjust lighting, etc., to the same level of precision typical of in-lab studies. Nearly all this work has used `Webgazer.js` — most in combination with jsPsych, but some with Gorilla or hand-built integrations.

Results to date have been encouraging. Semmelmann and Weigelt (2018) found data quality was reasonable for fixation location and saccades in fixation, smooth pursuit, and free-viewing tasks, though data collected online through a crowdsourcing platform was slightly more variable and timing was somewhat delayed compared to data collected in the lab. Several researchers successfully replicated well-known findings from the sentence-processing literature involving predictive looks (Degen, Kursat, & Leigh, 2021; Prystauka, Altmann, & Rothman, 2023; Van der Cruyssen et al., 2023; Vos, Minor, & Ramchand, 2022). Yang and Krajbich (2021) successfully replicated a well-established link between value-based decision-making and eye gaze (see also Van der Cruyssen et al., 2023).

While promising, there are some salient limitations. First, many of the studies report effects that are smaller or later than what had been previously observed in the lab (Degen et al., 2021; Slim & Hartsuiker, 2022; Van der Cruyssen et al., 2023). Importantly, accurate timing on a web browser is not trivial (De Leeuw & Motz, 2016; Passell et al., 2021), and subtle programming choices can significantly affect the accuracy of `WebGazer.js` timing (Yang & Krajbich, 2021).[1] Since most of the prior work did not address these timing issues, it is not clear how many of the reported lab/web differences would resolve.

Second, prior work has focused on studies with relatively coarse-grained regions of interest (but see Semmelmann & Weigelt, 2018), dividing the screen either in half or in quadrants. This is particularly salient with (Prystauka et al., 2023), which simplified (Altmann & Kamide, 1999)'s design so that regions of interest are quadrants rather than the finer-grained ROIs used in the original. Certainly, webcam eyetracking will not be as spatially fine-grained as an infrared eyetracker, but we do not yet have a good sense of the limits.

Finally, the prior work has focused on a relatively limited range of methods. Different paradigms have different technical requirements and analyze the results differently. As a

[1] See also discussion at https://github.com/jspsych/jsPsych/discussions/1892

result, the breadth of utility of webcam eye-tracking is unclear.

Present work

In order to validate the online eye-tracking methodology, with the particular configuration known to have the greatest temporal precision, `jsPsych` and a modification of `Webgazer`, we set out to reproduce five previously published studies representing a variety of questions, topics, and paradigms. The goal was to examine the strengths and weaknesses of webcam eye-tracking for common paradigms in cognitive science, across a broad range of research areas.

Selection of Studies

Studies with large effect sizes and which are known to replicate are ideal targets for further replication; otherwise, it can be difficult to distinguish a failure of the method from a failure of the original study to replicate. In practice, replications (successful or otherwise) have only been reported for a small number of studies, so we ultimately included some studies with unknown replicability. We addressed this in several ways. First, replicating five very different studies from different research traditions decreases our reliance on any one study. Second, we include several ‘’sanity check” analyses, such as the correlation between calibration accuracy and effect size. If the effect is real but there is noise from low-accuracy eye-tracking, this correlation should be substantial. Third, for two of the studies, we had comparison data collected in-lab either using `jsPsych` or a more traditional eyetracker technology, allowing us to directly assess the impact of differences in subject population and equipment/setting.

We chose five high-impact eye-tracking studies involving adult subjects (for an investigation of WebGazer's validity for developmental research, see Steffan et al., 2024 for a comparison of remote WebGazer and in-lab anticipatory looking effects in

Table 1

Studies selected for replication attempts. Citation counts based on Google Scholar (May 2024).

Citation	Topic Area	Paradigm	Citations
Altmann & Kamide, 1999	Psycholinguistics	Natural Scenes	2,130
Johansson & Johansson, 2014	Memory	Four Quadrants	259
Manns, Stark, & Squire, 2000	Memory	Two Halves	134
Snedeker & Trueswell, 2004	Psycholinguistics	Four Quadrants	487
Shimojo et al., 2003	Decision Making	Two Halves	1,146

18-27-month-old participants). Our goal was to include experiments from a range 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 "naturalistic" scenes). As noted above, we had a preference for well-established findings that are known to replicate, though for sake of diversity this was not always possible. Table 1 provides an overview of the five studies we selected.

General Methods

Participants

Participants completed the experiment remotely and were recruited through the Prolific platform. In order to have access to the experiment, participants had to meet the following criteria: 18 years of age or older, fluency in English, and access to a webcam. All participants provided informed consent. The online studies were approved by the Vassar College Institutional Review Board.

In addition, an in-lab replication was conducted for Experiment 1. Information about the sample is given in the Experiment 1 Method sections. This study was approved by the

Institutional Review Board at Boston College.

In order to have adequate statistical power and precision, we aimed for 2.5x the sample size of the original experiment, following the heuristic of Simonsohn (Simonsohn, 2015). In Experiment 5, the original sample size was so small that we opted to collect 5x the number of participants to increase precision. Because of budget and time constraints we were unable to replace the data for subjects who were excluded or whose data was missing due to technical failures.

Equipment

We used a fork of the `webgazer.js` library for webcam eye-tracking (Papoutsaki et al., 2016), implemented in `jsPsych`, a Javascript library for running behavioral experiments in a web browser (de Leeuw, 2015). Our fork included changes to `webgazer.js` in order to improve data quality for experiments in which the precise timing of stimulus onsets is relevant. Specifically, we implemented a polling mode so that gaze predictions could be requested at a regular interval, which improved the sampling rate considerably in informal testing. This modification is similar to what Yang and Krajbich (2021) reported improved the sampling rate in their study of webgazer. We also adjusted the mechanism for recording time stamps of each gaze prediction, so that the time stamp reported by webgazer is based on when the video frame is received and not when the computation of the gaze point is finished.

Eye-tracking Calibration and Validation

When participants began the experiment, they were notified the webcam would be used for eye tracking but no video would be saved. They were asked to remove glasses if possible, close any other tabs or apps, turn off notifications, and make sure their face was lit from the front. The webcam's view of the participant popped up on the screen, and participants were asked to center their face in the box and keep their head still. The experiment window then expanded to full screen, and participants began the eye-tracking calibration.

During the calibration, dots appeared on the screen one at a time in different locations, and the participants had to fixate them and click on each one. Once they clicked on a dot, it would disappear and a new one would appear in a different location on the screen. The locations of calibration dots were specific to each experiment (details below) and appeared in the areas of the screen where the visual stimuli would appear during the main task in order to ensure that eye movements were accurately recorded in the relevant regions of interest. After the calibration was completed, the validation began. Participants were asked to go through the same steps as the calibration, except that they only fixated the dots as they appeared in different locations on the screen. If accuracy on the validation was too low (fewer than 50% of looks landed within a 200 px radius of the validation points), participants were given an opportunity to re-start the calibration and validation steps.

Pre-registration

These data were collected within the context of an undergraduate research methods course. Groups of students (co-authors) designed and programmed experiments in jsPsych, pre-registered their planned analyses, and collected data through Prolific under the supervision of the first author. The OSF repositories associated with these experiments are linked in the methods sections of each individual study. Note that in the current paper we expand on those pre-registered analyses (e.g., including analyses of the calibration quality). All analysis code underlying this paper can be found in the Github repository: https://github.com/jodeleeuw/219-2021-eyetracking-analysis

Data Pre-processing

We used R (Version 4.2.1; R Core Team, 2021) and the R-packages *afex* (Version 1.3.0; Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2021), *broom.mixed* (Version 0.2.9.4; Bolker & Robinson, 2020), *dplyr* (Version 1.1.4; Wickham, François, Henry, & Müller, 2021), *forcats* (Version 1.0.0; Wickham, 2021a), *ggplot2* (Version 3.4.4; Wickham, 2016), *jsonlite* (Version 1.8.8; Ooms, 2014), *lme4* (Version 1.1.35.1; Bates, Mächler, Bolker, & Walker, 2015), *lmerTest* (Version 3.1.3; Kuznetsova, Brockhoff, & Christensen, 2017), *Matrix* (Version 1.6.5; Bates & Maechler, 2021), *papaja* (Version 0.1.2; Aust & Barth, 2020), *readr* (Version 2.1.5; Wickham & Hester, 2020), *shiny* (Version 1.8.0; Chang et al., 2021), *stringr* (Version 1.5.1; Wickham, 2019), *tidyr* (Version 1.3.1; Wickham, 2021b), and *tinylabels* (Version 0.2.4; Barth, 2022) for all our analyses.

Experiment 1a

The first study was a replication attempt of Altmann and Kamide (1999). Altmann and Kamide used the visual world eye-tracking paradigm (Tanenhaus et al., 1995) to show that meanings of verbs rapidly constrain the set of potential subsequent referents in sentence processing. For example, when looking at the display in Figure 1 and listening to a sentence like "The boy will eat the…," participants are more likely to look at the cake than when they hear "The boy will move the…," in which case they tend to look at the train, presumably because cakes are edible and trains are not. Semantic information available at the verb is used to anticipate upcoming linguistic input.

Method

All stimuli, experiment scripts, data, analysis scripts, and a pre-registration are available on the Open Science Framework at https://osf.io/s82kz.

Participants. Sixty participants were paid $2.60 for their participation. Our sample size of participants was determined by the total run time of our experiment, ~10 minutes, and the allotted funding from the Vassar College Cognitive Science Department. From this information, we calculated a reasonable number of participants we could afford to compensate on Prolific. Note that the sample size of the original study was 24. For unknown reasons, 2 of the subjects' results were not recorded, so in the analysis, we worked with data collected from 58 participants.

Materials and Design. The visual stimuli were created through Canva and depicted an agent accompanied by four to five objects in the scene (see Figure 1). On critical trials, participants heard one of two sentences associated with the scene. In the restrictive condition, the sentence (e.g., "The boy will eat the cake") contained a verb (e.g., "eat") which restricts the set of possible subsequent referents (e.g., to edible things). Only the target object (e.g., the cake) was semantically consistent with the verb's meaning. In the non-restrictive condition, the sentence (e.g., "The boy will move the cake") contained a verb (e.g., "move") which does not restrict the set of possible subsequent referents. The target object (e.g., the cake) as well as the distractor objects (e.g., the train, the ball, etc.) were semantically consistent with the verb's meaning. Both sentences were compatible with the scene, such that the correct keyboard response for the critical trials was "yes." Filler trials consisted of scenes that also contained an agent surrounded by objects as in the critical trials, but corresponding sentences named an object that was not present in the scene. The correct keyboard response for the filler trials was "no."

Each participant was presented with 16 critical trials (eight in the restrictive condition, eight in the non-restrictive condition) and 16 fillers for a total of 32 trials. The order of trials and the assignment of critical scene to condition was random on a subject-by-subject basis.

Procedure. The task began with a 9-point eye-tracker calibration and validation (Figure 2). During the experiment, the participants were simultaneously presented with a

Figure 1. Example trial from Experiment 1. Participants would hear a sentence (e.g., “The boy will eat the cake”) and respond according to whether the sentence matched the picture.

visual image and a corresponding audio recording of a spoken sentence. Participants had to input a keyboard response indicating “yes” or “no” as to whether the sentence they heard was feasible given the visual image. There were two practice trials to ensure that participants understood the instructions before they undertook the main portion of the experiment. Participants’ reaction times, keyboard responses, and looks to objects in the scene were recorded for each trial.

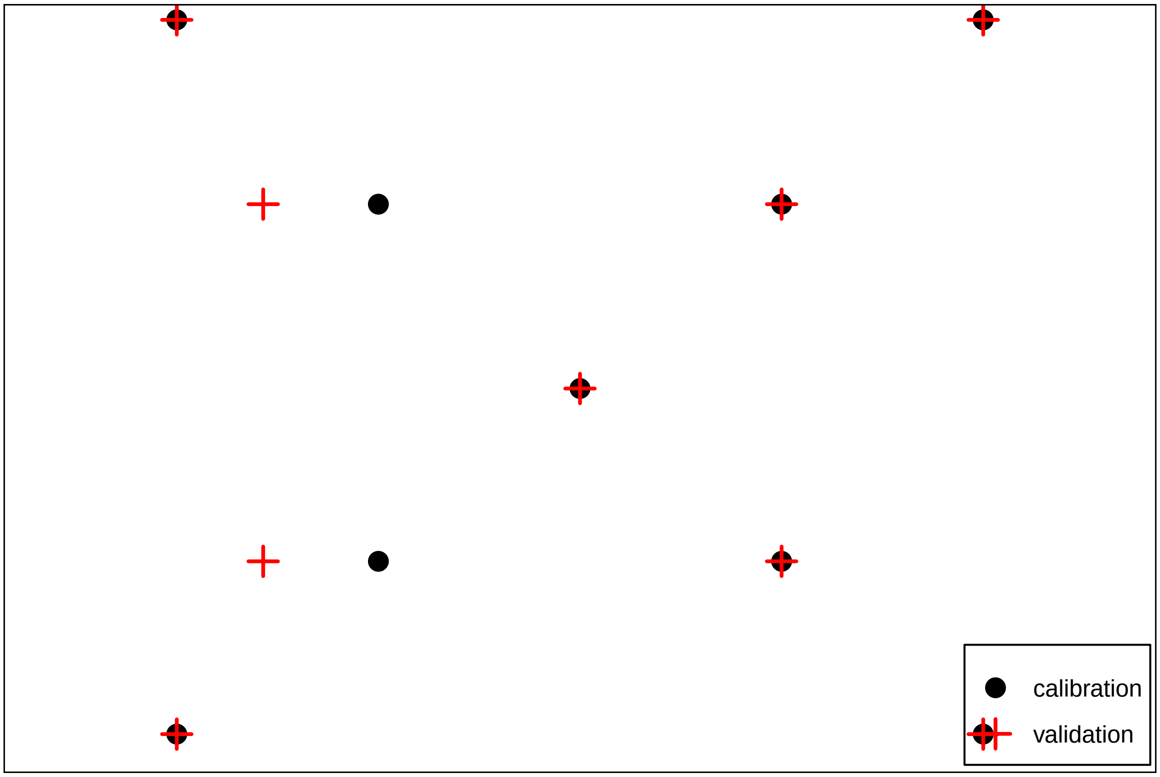


Figure 2. Calibration and validation point locations for Experiment 1. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.

Results

Looks to the objects in the scene were time-locked to the onset of the verb, the offset of the verb, onset of the post-verbal determiner, and onset of the target noun. ROIs were defined by creating boxes around each object in the scene. The size of each box was determined by taking the height and width of the given object and adding 20 pixels of padding. Each scene contained an agent region, a target region, and three or four distractor regions.

Cumulative Fixation Probabilities. For each sentence, the target time window began at the onset of the verb and ended 2000 milliseconds later. This window was then divided into 50-ms bins; for each participant and each trial, we recorded whether each object was fixated during the 50-ms bin. Collapsing over trials and participants, and averaging across distractors, we calculated the cumulative probability of fixation, shown in Figure 4, Panel (b).

Pre-noun fixations. In our first two analyses, we asked whether participants looked more to the target than to the distractor during the predictive time window, given that the verb is restricting. The first model tested whether there were more fixations to the target object than to the distractor in the time window before the onset of the target noun. We ran a regression model predicting the cumulative fixation probability in the last 50-ms bin before noun onset from the verb condition (restricting = 1 vs. non-restricting = 0), object type (target = 1 vs. distractor = 0), and their interaction, along with random effects for participants and images (with no covariance between random effects because the model cannot converge with full covariance matrix)^[`lme4` syntax: `lmer_alt(probability ~ object_type*verb_condition + (object_type*verb_condition || subject) + (object_type*verb_condition || scene)`. There were no significant effects, although the critical interaction was in the expected direction (b = 0.05, SE = 0.03, p=0.15).

Pre-verb-offset fixations. Altmann and Kamide tested a second model, aligning the predictive time window with the offset of the verb rather than the onset of the noun as above. When we did the same^[`lme4` syntax: `lmer_alt(probability ~ object_type*verb_condition + (object_type*verb_condition || subject) + (object_type*verb_condition || scene)`, we again saw that the critical interaction is not significant but numerically in the expected direction (b = 0.05, SE = 0.03, p=0.20).

First target fixations after verb. Finally, we addressed whether participants looked to the target faster in the restrictive vs. the non-restrictive condition, starting after the onset of the verb. On average, participants looked to the target 335 ms after the noun onset in the restrictive condition James, Minnihan, & Watson (2023) and 435 ms after the noun onset in the non-restrictive condition James et al. (2023). Thus, first fixations were not only delayed relative to those in the previous studies compared here, but also showed a smaller difference between conditions.

We ran a regression model predicting the timing of the first fixation to the target object, relative to the onset of the noun, with verb condition as a predictor, mean-centered verb duration as a covariate, and random intercepts and condition slopes for participants and scenes^[`lme4` syntax: `lmer_alt(time ~ verb_condition*verb_duration + (verb_condition || subject) + (verb_condition || scene)`. There were no significant effects; participants looked sooner at the target in the restrictive condition, while accounting for verb duration and its interaction with condition, but this was not a statistically significant effect (b = -121.60, SE = 85.43, p=0.17).

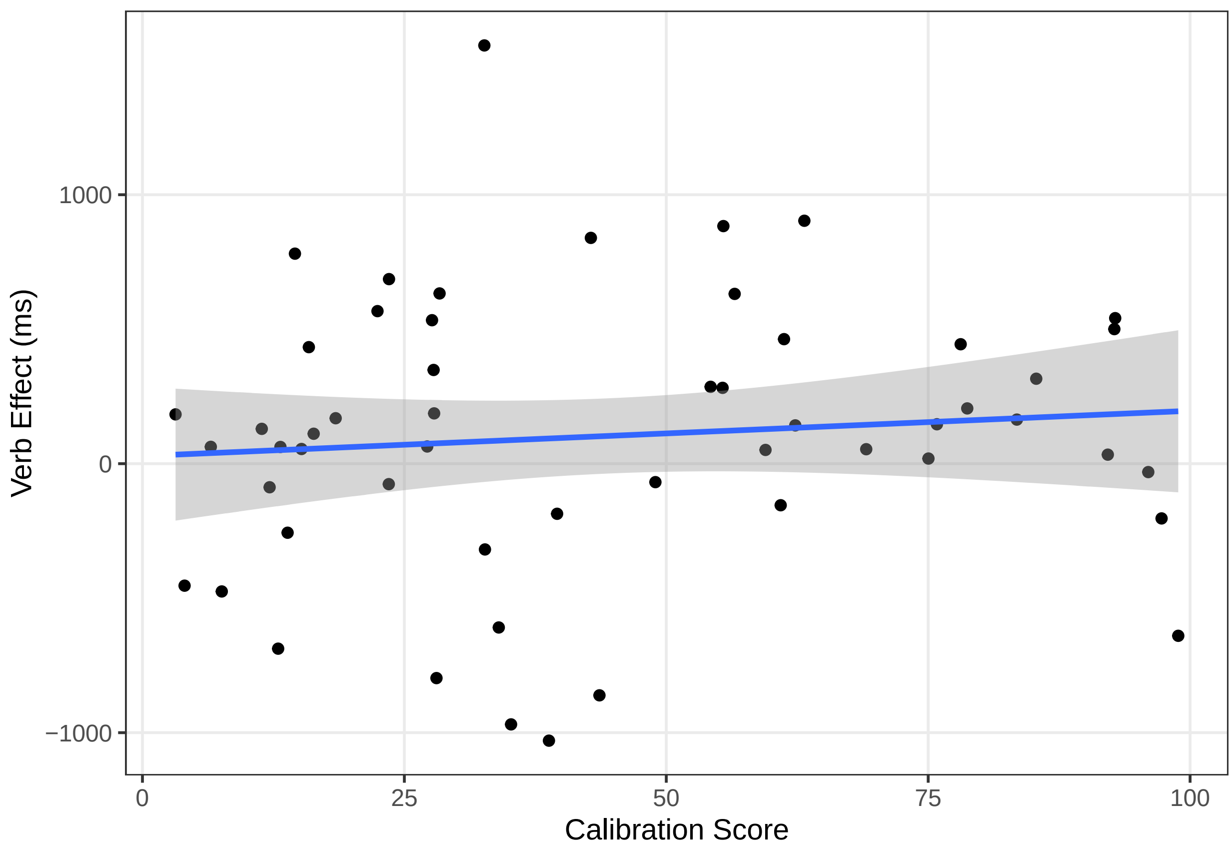


Figure 3. Calibration scores plotted against the verb effect on first fixations (the latency of the first target fixation in the non-restricting verb condition minus that in the restricting verb condition).

Calibration. Participants' calibration quality was measured as the mean percentage of fixations that landed within 200 pixels of the calibration point. Calibration quality varied widely, ranging from 3.16% to 98.87%.

We tested whether a participant's calibration quality was correlated with their effect size. There were three effects of interest: the verb-by-object interaction in predicting fixation probabilities, both in the (1) pre-noun-onset and (2) pre-verb-offset windows (calculated as the difference in target-over-distractor preference between verb conditions), and (3) the effect of verb on the timing of the first target fixation (calculated as the difference in target latency between verb conditions). Across the three effects of interest, calibration quality was not significantly correlated (Effect 1: Pearson's r = 0.03, p = 0.83, Effect 2: Pearson's r = -0.05, p = 0.73, Effect 3: Pearson's r = 0.10, p = 0.49. Figure 3 plots the relation between calibration scores and Effect 3. The figure makes clear that there is a large proportion of participants with calibration scores under 50%.

Re-analysis After Exclusions. We re-analyzed the data after excluding participants with calibration scores under 50% (N = 35). The following analyses thus included 22 participants.

We tested the same three models under these more aggressive exclusion criteria. The first two models, comparing target and distractor fixations in the predictive window, produced very similar results; the critical interaction was not statistically significant (Pre-noun-onset window: b = 0.07, SE = 0.06, p=0.23; Pre-verb-offset window: b = 0.05, SE = 0.05, p=0.28). However, the final model, which tested the effect of verb condition on the timing of fixations to the target, yielded a statistically significant result, unlike in the previous set of analyses with all participants included (b = -193.35, SE = 96.33, p=0.05).

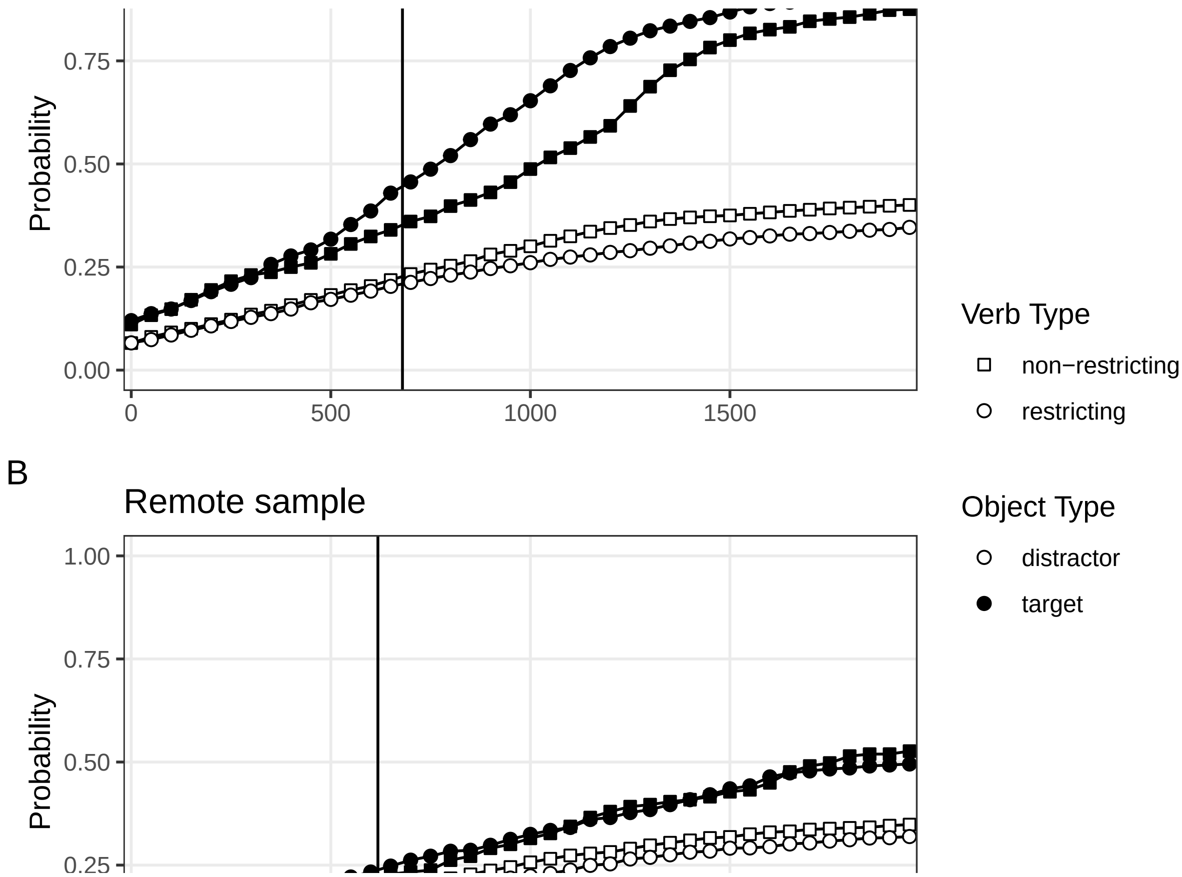


Figure 4. Cumulative probability of fixating distractor and target objects across conditions over time, with 0 ms aligned to the verb onset time. The vertical line marks the mean noun onset time across trials and conditions.

Discussion

Across three different tests of the hypothesis that listeners will use verb semantics to anticipate the upcoming referent, we found results that were numerically consistent with the hypothesis but not statistically significant. A comparison to published data (James et al., 2023) demonstrates that looks to the objects in the scene (relative to background or off-screen looks) were depressed across conditions and objects. After eliminating participants with validation accuracy under 50% and/or 10% or fewer fixations to any ROIs, we were left with only 22 of the original 60 participants. Analyses on this subset resulted in one effect reaching statistical significance (shorter target fixation latencies in the restricting vs. non-restricting verb condition). Given that nearly two-thirds of participants had poor data quality, we ran a follow-up webcam study in a lab setting order to isolate the potential causes.

Experiment 1b

Experiment 1b tested whether the failure to replicate significant condition effects in Experiment 1a was due to features of conducting the study remotely (i.e. varied experimental settings and apparatuses, lower compliance) rather than webcam-based eye-tracking or `webgazer` *per se*. Thus, Experiment 1b took place in a lab setting with undergraduate participants but otherwise used the same Method as Experiment 1a.

Method

Participants. Forty-nine participants completed the study in a lab setting. They were recruited via the Boston College subject pool. Participants needed to be 18 years of age or older and native speakers of English.

Materials and Design. Materials were identical to those in Experiment 1a.

Procedure. After being greeted by the experimenter and completing the informed consent form, participants followed on-screen prompts to complete the study, including calibration, as described in the Experiment 1a Procedure.

Apparatus. HELP

Results

Cumulative Fixation Probabilities. For each sentence, the target time window began at the onset of the verb and ended 2000 milliseconds later. This window was then divided into 50-ms bins; for each participant and each trial, we recorded whether each object was fixated during the 50-ms bin. Collapsing over trials and participants, and averaging across distractors, we calculated the cumulative probability of fixation, shown in Figure 6, Panel (b). The results from Experiment 1a are copied here for ease of comparison (Panel a).

Pre-noun fixations. In line with the previous analyses of the remote data, we asked whether participants looked more to the target than to the distractor object during the predictive time window, depending upon the verb condition. The first model constrained the predictive window to the time before the onset of the target noun. We ran a regression model predicting the cumulative fixation probability in the last 50-ms bin before noun onset from the verb condition (restricting = 1 vs. non-restricting = 0), object type (target = 1 vs. distractor = 0), and their interaction, along with random effects for participants and scenes (with no covariance between random effects because the model cannot converge with full covariance matrix). Unlike the remote sample, there was a significant main effect of object type such that participants were more likely to be looking at the target than the distracter object during this time window (b = 0.10, SE = 0.03, p=0.01). Also unlike the remote sample, the critical interaction was not in the expected direction, although it was also not statistically significant (b = -0.05, SE = 0.05, p=0.25).

Pre-verb-offset fixations. Altmann & Kamide tested a second model, aligning the predictive time window with the offset of the verb rather than the onset of the noun as above. When we did the same, we again saw that the critical interaction is not significant nor in the expected direction (b = -0.06, SE = 0.04, p=0.17).

First target fixations after verb. Finally, we addressed whether participants looked to the target faster in the restrictive vs. the non-restrictive condition, starting after the onset of the verb. On average, participants looked to the target 405 ms after the noun onset in the restrictive condition and 399 ms after the noun onset in the non-restrictive condition. As in the remote sample, the latencies are overall slower than in results published by Altmann & Kamide (1999) and James et al. (2023). Unlike in the remote sample, however, the difference is in the unexpected direction, such that participants looked to the target faster in the non-restrictive condition.

We ran a regression model predicting the timing of the first fixation to the target object, relative to the onset of the noun, with verb condition as a predictor, mean-centered verb duration as a covariate, and random intercepts and condition slopes for participants and scenes. There were no significant effects; results revealed that the paradoxical advantage in the non-restrictive condition was not statistically significant (b = 2.88, SE = 117.68, p=0.98). Effects of verb duration and its interaction with condition were also not statistically significant (duration: b = -0.58, SE = 0.54, p=0.30; interaction: b = -0.60, SE = 0.96, p=0.53)

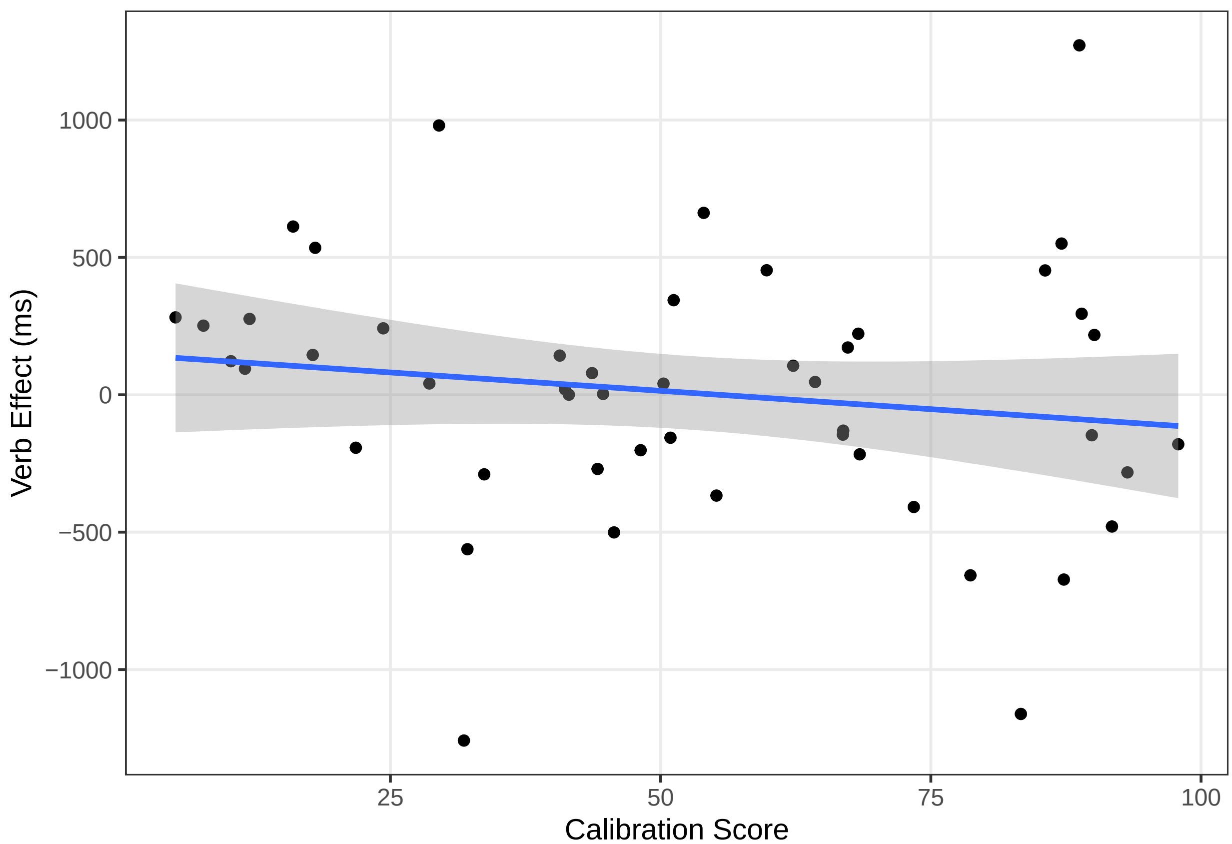


Figure 5. Calibration scores plotted against the verb effect, operationalized as the latency of the first target fixation in the non-restricting verb condition minus that in the restricting verb condition.

Calibration. As before, participants' calibration quality was measured as the mean percentage of fixations that landed within 200 pixels of the calibration point. Calibration quality ranged from 5.13% to 97.89%. Across the three condition effects of interest, calibration quality was not significantly correlated (pre-noun: Pearson's r = -0.24, p = 0.11; pre-verb-offset: Pearson's r = -0.19, p = 0.20; first fixation: Pearson's r = -0.16, p = 0.29. Figure 3 plots the relation between calibration scores and Effect 3.

Re-analysis After Exclusions. Again, we excluded participants with validation accuracy under 50%, which eliminated 23 participants. The following analyses included 26 participants.

Across all three models, results were in line with analyses using the minimal exclusion criteria; none of the critical effects were statistically significant, nor were they in the expected direction (Pre-noun-onset window, verb x object interaction: b = -0.08, SE = 0.06, p=0.15; Pre-verb-offset window, verb x object interaction: b = -0.07, SE = 0.05, p=0.16; Verb effect on first target fixation: b = 6.40, SE = 97.92, p=0.95).

Discussion

Overall, the results of Experiment 1 paint a sobering picture of web-based eye-tracking. In Experiment 1a, results from the remote samplewere in the expected direction but effects were smaller and delayed relative to previous work and failed to reach statistical significance. To test whether this could be explained by the variability in experimental settings across participants, we replicated our procedure in a lab setting. Surprisingly, the results in the in-lab study were less aligned with previous work; critical effects were not significant nor were they in the expected direction numerically. This suggests the problem was not with running the study online or due to working with a more diverse population via Prolific.

Taken together, the instability of these effects might suggest that this paradigm is not

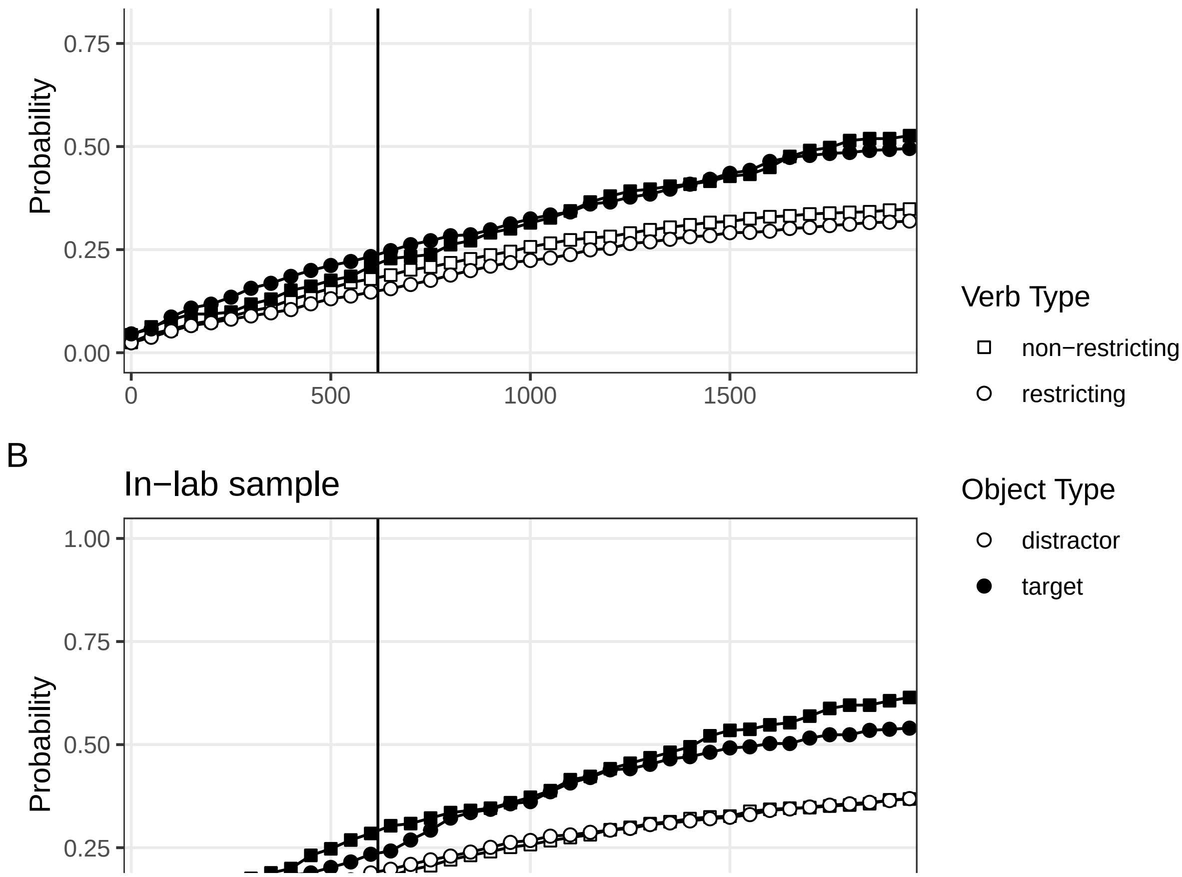


Figure 6. Cumulative probability of fixating distractor and target objects across conditions over time, with 0 ms aligned to the verb onset time. The vertical line marks the mean noun onset time across trials and conditions.

well-suited for webcam-based eye-tracking. Notably, the ROIs were tightly drawn around the five to six objects in each scene (drawing larger ROIs in these scenes would have led to overlapping objects) and thus, analyses were unforgiving of inaccurate calibration. Further evidence comes from a recent `webgazer` study that successfully replicated a modified version of the Altmann and Kamide study with only four objects, each in separate quadrants, allowing for larger, more distinct ROIs (Prystauka et al., 2023).

The next four experiments test paradigms with more generous ROIs.

Experiment 2

The second study was a replication attempt of Johansson and Johansson (2014), which examined how visuospatial information is integrated into memory for objects. They found that, during memory retrieval, learners spontaneously look to blank screen locations where pictures were located during encoding (see Spivey & Geng, 2001) and that this spatial reinstatement facilitates retrieval of the picture.

Method

All stimuli, experiment scripts, data, analysis scripts, and a pre-registration are available on the Open Science Framework at https://osf.io/xezfu/.

Participants. Sixty participants were paid for their participation. The sample size was motivated in part by budget constraints, but was nonetheless 2.5x larger than the original sample size of 24). Data from one participant were not properly recorded due to unknown technical issues, so data from 59 participants were included in all analyses to follow.

Materials and Design. The experiment consisted of two blocks each composed of an encoding phase and a recall phase. The two blocks differed by whether the recall phase was in the free-viewing or fixed-viewing condition, as described in the Procedure. Participants were randomly assigned to see the fixed-viewing or free-viewing block first. During each encoding phase, participants saw a grid indicating the four quadrants of the screen. Each quadrant contained six cartoon images of items belonging to the same category. The four categories were humanoids, household objects, animals, and methods of transportation (see Figure 7). Different images were used in each block; there were 48 unique images total across the experiment.

Each recall phase presented participants with a blank screen with a central fixation cross as they listened to true/false statements testing their recall of the previous grid. Each statement 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 in each block, such that each object was the subject of both an intraobject and an intraobject statement; there were 96 unique statements total across the experiment.

Figure 7. Example trial from Experiment 2.

Procedure. The task began with a 9-point eye-tracker calibration and validation (Figure 8).

Participants received instructions that included an example grid and an explicit request that they not use any tools to help them during encoding. Participants then began their first encoding phase. Each of the four quadrants was presented one at a time. First, a list of the items in the quadrant was shown, then the pictures of items were displayed in the quadrant. For each item, participants used their arrow keys to indicate whether the object was facing left or right. After the participant identified the direction of each item,

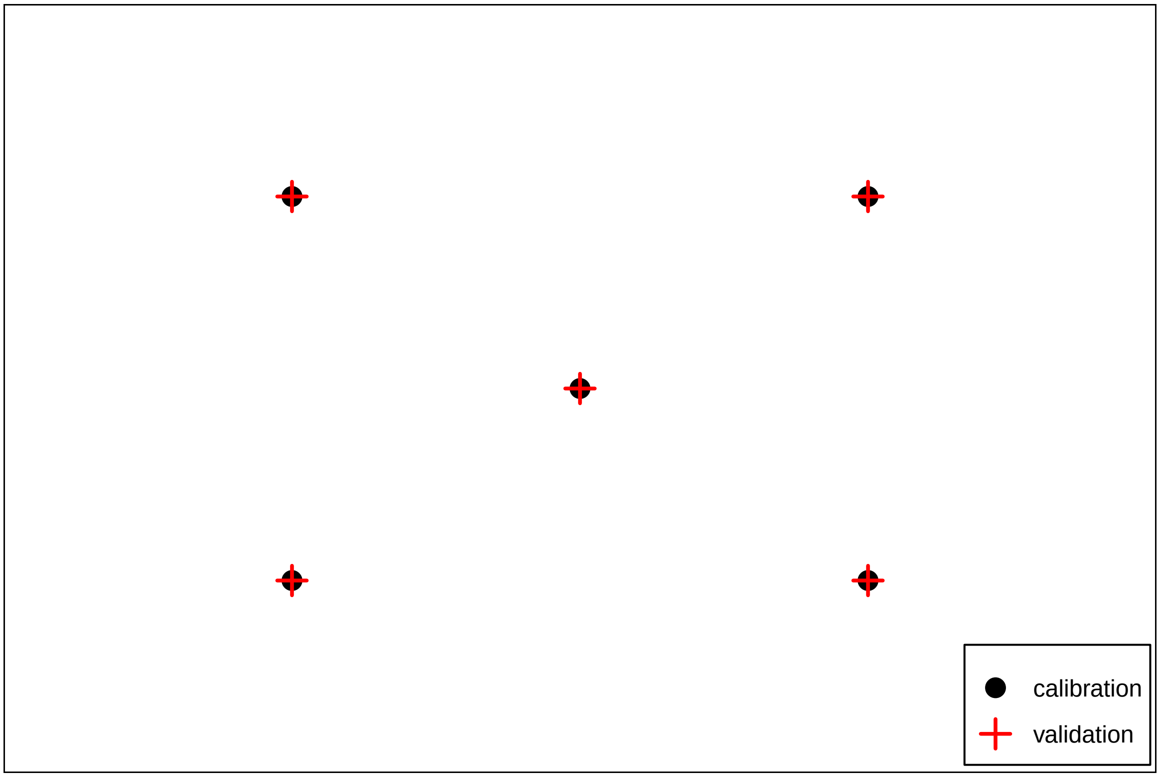


Figure 8. Calibration and validation point locations for Experiment 2. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.(In this experiment all the same locations were used for both calibration and validation.)

they would have an additional 30 seconds to encode the name and orientation of each item in the quadrant. Finally, after all four quadrants were presented, participants were shown the full grid of 24 items and had 60 seconds to further encode the name and orientation of each item.

Participants then entered the first recall phase, in which they listened to the 48 statements and responded by pressing the 'F' key for false statements and 'T' for true ones. While listening to these statements, in the free-viewing block, participants saw a blank screen and were allowed to freely gaze around the screen. During the fixed-viewing block, participants were asked to fixate a small cross in the center of the screen throughout the recall phase. In both cases, the mouse was obscured from the screen.

Participants then proceeded to the second encoding and recall phases as described above. After completing both encoding-recall blocks, participants were asked to answer a few survey questions (such as whether they wore glasses or encountered any distractions).

The primary methodological difference between this replication and Johansson and Johansson's study was that the original study included two additional viewing conditions that were omitted from this replication due to time constraints. In those two conditions, participant were prompted to look to a specific quadrant (rather than free viewing or central fixation) which either matched or mismatched the original location of the to-be-remembered item.

Results

Eye-gaze. Looks during the retrieval period were categorized as belonging to one of four quadrants based on the x,y coordinates. The critical quadrant was the one in which the to-be-retrieved object had been previously located during encoding. The other three quadrants were labeled "first", "second," and "third" depending upon the location of the critical quadrant (e.g., when the critical quadrant was in the top left, the "first" quadrant was the top right quadrant, but when the critical quadrant was in the top right, "first" corresponded to bottom right, etc.). In both the fixed- and free-viewing condition, participants directed a larger proportion of looks to the critical quadrant (see Figure 9). This bias appeared larger in the free-viewing condition, suggesting that the manipulation was somewhat effective.

The proportions of looks across quadrants in the free-viewing condition were analyzed using a linear mixed-effects model with quadrant as the predictor (critical as the reference level). The model included random intercepts and slopes for participants[2]. Proportions of

[2] `lme4` syntax: `lmer(proportion ~ quadrant + (1+quadrant|subject_id))`. Among other limitations, this approach violates the independence assumptions of the linear model because looks to the four

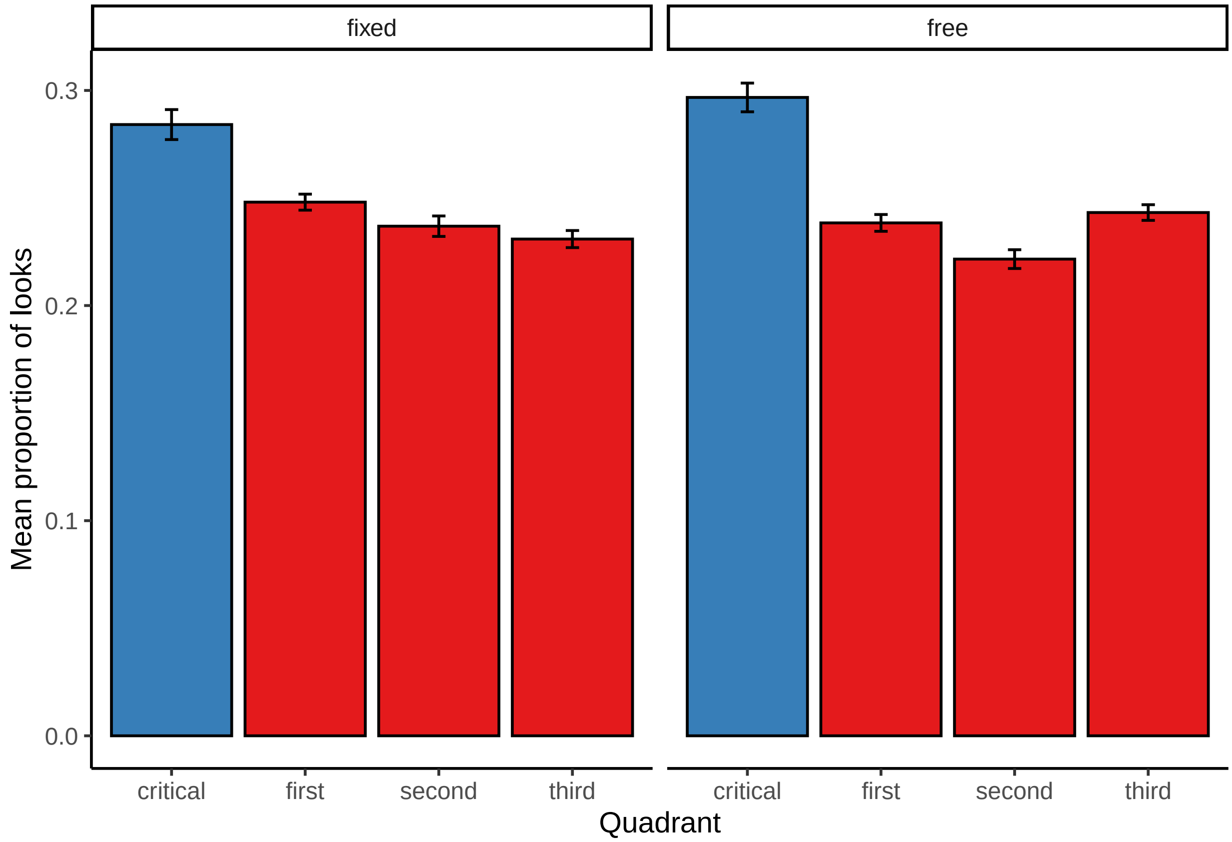


Figure 9. Proportion of eye-gaze to critical quadrant and other three quadrants during memory retrieval in a) fixed and b) free viewing conditions.

looks were significantly higher for the critical quadrant compared to the other three (first: b = -0.06, SE = 0.01, p<0.001, second: b = -0.08, SE = 0.01, p<0.001, third: b = -0.05, SE = 0.01, p<0.001).

Response Time and Accuracy. Participants' response times and accuracies on memory questions are summarized in Figure 10. Both dependent variables were analyzed with linear mixed-effects model with relation type (interobject = -0.5, intraobject=0.5) and viewing_condition (fixed = -0.5, free=0.5) and their interaction as the predictors. The

locations are not independent. This analysis was chosen because it is analogous to the ANOVA analysis conducted in the original paper.

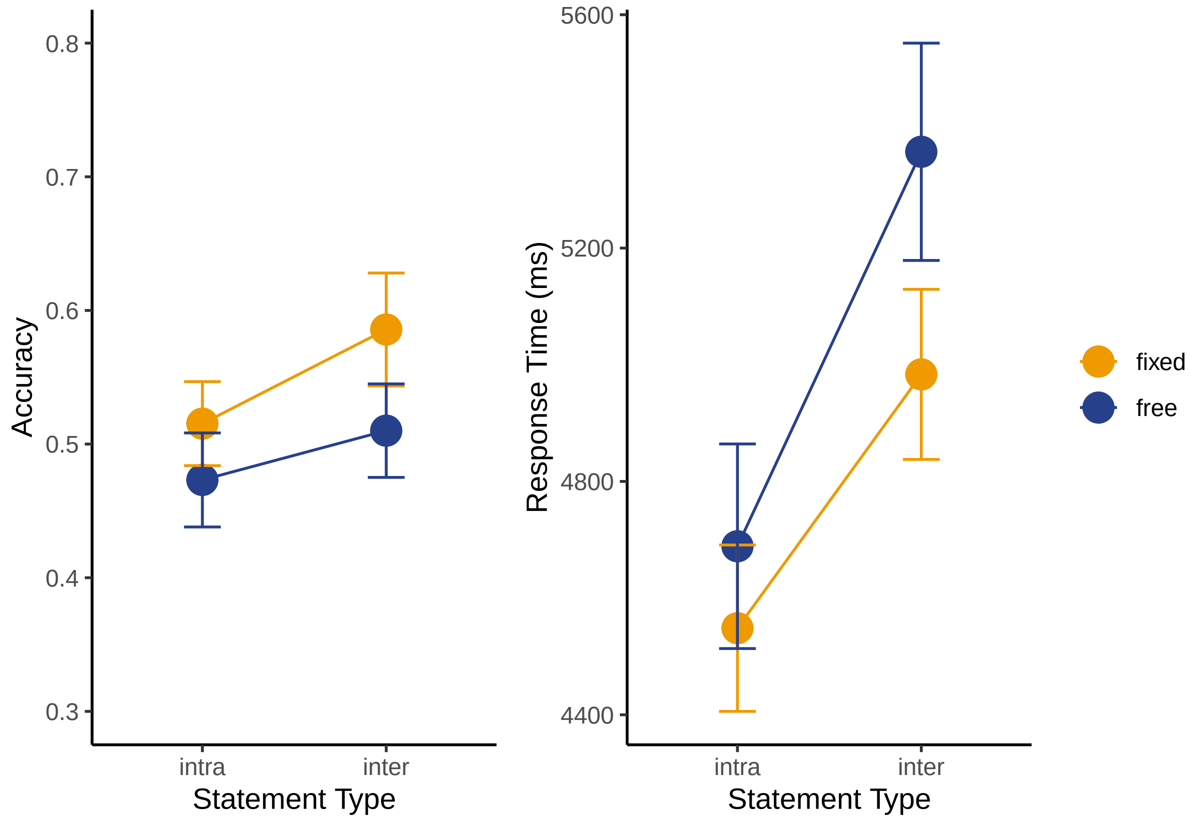


Figure 10. Accuracy and response times during memory retrieval.

model included random intercepts for participants[3]. Accuracy did not differ significantly between interobject and intraobject questions (b = -0.05, SE = 0.03, p=0.05). Participants were less accurate in the free viewing condition than the fixed condition (b = -0.06, SE = 0.03, p=0.03). Response times were slower for interobject (e.g., "The train is to the right of the taxi.") than intraobject (e.g., "The train is facing right.") questions (b = -555.60, SE = 105.24, p<0.001). Response times were slower in the free viewing condition than the fixed condition (b = 260.98, SE = 105.24, p<0.001). The interaction was not a significant predictor for response times or accuracy. These behavioral results are inconsistent with the original findings.

One possibility is that the in-lab participants in Johansson and Johansson (2014)

[3] `lme4` syntax: `lmer(DV ~ relation_type*viewing_condition + (1|subject_id))`

were much more compliant with the instruction to keep their gaze on central fixation (though these data are not reported in the original paper). When analyzing results from the subset of participants (N = 25) who were most compliant during the fixed-viewing block (at least 25% of their looks fell within 20% of the center of the display), the viewing condition effects and the interactions were not signficant. Given the smaller sample size we do not interpret these results further.

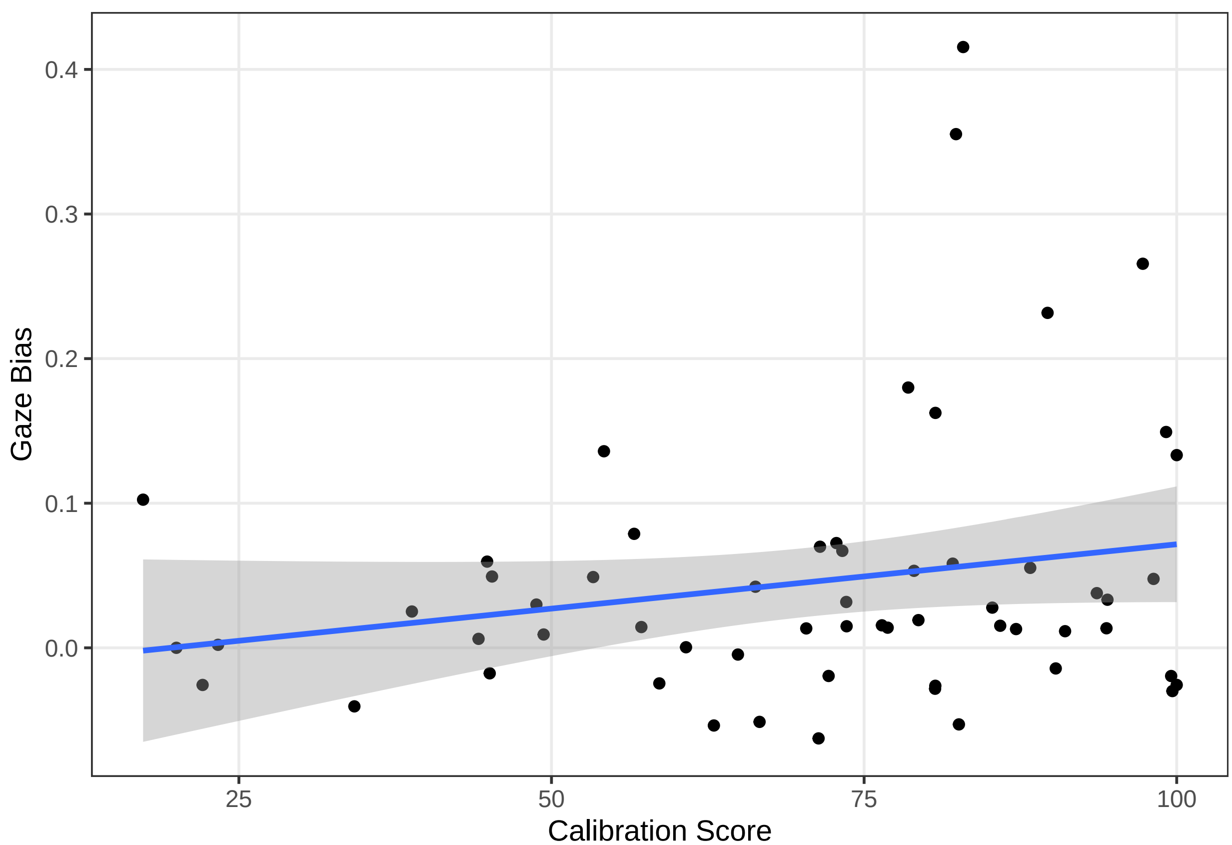


Figure 11. Calibration scores plotted against gaze bias (the difference between the proportion of looks to the critical quadrant minus the average proportion of looks to the average of the other three quadrants).

Calibration. Participants' calibration quality was measured as the mean percentage of fixations that landed within 200 pixels of the calibration point, averaging initial calibration (or re-calibration for participants who repeated calibration) and calibration at the halfway point. Calibration scores varied substantially (between 17.34 and 100 %). The quality of a participant's calibration was not significantly correlated with the participant's effect size (*Pearson's* r= 0.21, p = 0.11) as measured by the difference between the proportion of looks to the critical quadrant minus the average proportion of looks to the average of the other three quadrants (see Figure 11).

Re-analysis After Exclusions. There were 12 participants whose calibration score was under 50%. When those participants were removed from the analyses, critical results were in line with the main analyses: there remains a clear gaze bias in the free-viewing condition (first: b = -0.06, SE = 0.01, p<0.001, second: b = -0.08, SE = 0.02, p<0.001, third: b = -0.06, SE = 0.01, p<0.001); viewing condition still did not interact with question type to predict either behavioral outcome.

Discussion

As in Johansson and Johansson (2014) and Spivey and Geng (2001), during memory retrieval, learners spontaneously look to blank screen locations where pictures were located during encoding, suggesting that visuospatial information is integrated into the memory for objects. However, we did not observe a memory benefit, in terms of speed or accuracy, of spatial reinstatement via gaze position during retrieval of the picture. We can speculate that this may be due to the fact that participants struggled to maintain their gaze fixed in the center in the fixed-viewing condition, such that the difference between the fixed- and free-viewing conditions was minimal. Crucially for the current purposes, the webcam-based eye-tracking measurements were successful in replicating the key eye-tracking results.

Experiment 3

The third study was a partial replication attempt of Manns, Stark, and Squire (2000). This experiment used the visual paired-comparison, which involves presenting a previously-viewed image and novel image together and measuring the proportion of time spent looking at each image. The expected pattern of results is that participants will look more at novel objects. They Manns et al. (2000) hypothesized that this pattern of behavior could be used to measure the strength of memories. If a viewer has a weak memory of the old image, then they may look at the old and new images roughly the same amount of time. They tested this in two ways. First, they showed participants a set of images, waited five minutes, and then paired those images with novel images. They found that participants spent more time (58.8% of total time) looking at the novel images. They then measured memory performance one day later and found that participants were more likely to recall images that they had spent less time looking at during the visual paired-comparison task the previous day.

Method

The stimuli, experimental code, and data and analysis scripts can be found on the Open Science Framework at https://osf.io/k63b9/. The pre-registration for the study can be found at https://osf.io/48jsv . We inadvertently did not create a formal pre-registration using the OSF registries tool, but this document contains the same information and is time stamped prior to the start of data collection.

Participants. Our pre-registered target was 50 participants. 51 participants completed the first day of the experiment and 48 completed the second day. Following Manns et al., we excluded 3 participants due to perfect performance on the recognition memory test because this prevents comparison of gaze data for recalled vs. non-recalled images. Our final sample size was 45 participants.

Materials and Design. Stimuli consisted of color photographs of common objects (e.g. an apple, a key, etc.). We selected 96 unique images from the stimulus set provided by Konkle, Brady, Alvarez, and Oliva (2010), which contains multiple objects from hundreds of unique categories. We selected images in 48 pairs from the same object category such that each critical object had a corresponding foil object during the recognition test (e.g. one red apple and one green apple). All images are of a single object on a white background.

Stimulus sets were created to present participants with 24 presentation trials and 24 test trials on Day 1, and 48 recognition trials on Day 2. Each presentation trial screen presented two identical images (e.g. red apple and red apple). The test trial screens presented each of the previously-seen (“old”) images paired with a new image (e.g. red apple and bicycle key). The recognition trial screens presented a single image that was either from the original set of 24 (seen twice across the first two phases) or was a corresponding foil for the original set (e.g. red apple or green apple). Thus, each participant was exposed to 72 unique images over the course of the experiment (the 24 foils for the “new” images in the test phase are never seen). Two full stimulus lists were created to counterbalance images across participants; the images that composed the “old” set for one list made up the “new” set for the list. The experimental design is visually depicted in Figure 13.

Procedure. The task began with a 7-point eye-tracker calibration (each point was presented 3 times in a random order) and validation with 3 points (each presented once). The point locations were designed to focus calibration on the center of the screen and the middle of the left and right halves of the screen (Figure 12).

The experiment was administered over the course of two consecutive days. The presentation phase and test phase occurred on the first day, while the recognition test occurred on the second day.

During the presentation phase, participants viewed 24 pairs of identical images. Each pair was presented for 5 seconds and an interval of 5 seconds elapsed before the next pair was shown. The order of the photographs was randomized and different for each participant. After completion of the presentation phase, participants were given a 5-minute break during which they could look away from the screen.

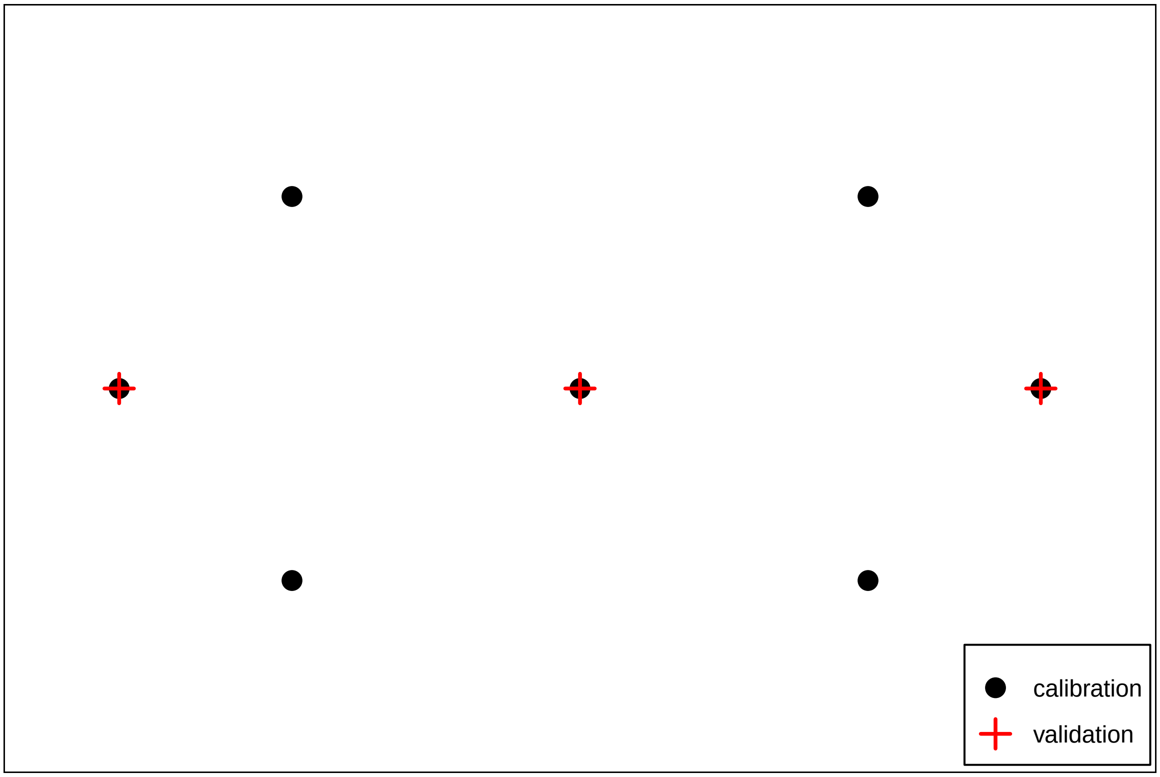


Figure 12. Calibration and validation point locations for Experiment 3. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.

After the break, they were prompted to complete the eye-tracking calibration again before beginning the test phase. During this phase, participants again viewed 24 pairs of photographs (old and new) with an interstimulus duration of 5 seconds.

Approximately 24 hours after completing the first session, with a leeway interval of 12 hours to accommodate busy schedules, participants were given the recognition test. Each image was shown on the screen for 1 second, followed by a 1 second interstimulus interval. Each photograph remained on the screen until the participants indicated whether or not they had seen it before by pressing ‘y’ for yes and ‘n’ for no. After they pressed one of the two keys, a prompt on the screen asked them to rate their confidence in their answer from 1 as a “pure guess” to 5 as “very sure.” by clicking on the corresponding number on the screen. No feedback on their responses was given during the test.

There were two modifications we made to the methods of the original experiment. As we were only replicating the declarative memory component of the original experiment, we did not have a "priming group." Therefore, we followed only the procedure for the "looking group." Additionally, for each section of the study, the stimuli were presented on a single screen instead of two screens due to the constraints of the online experiment format.

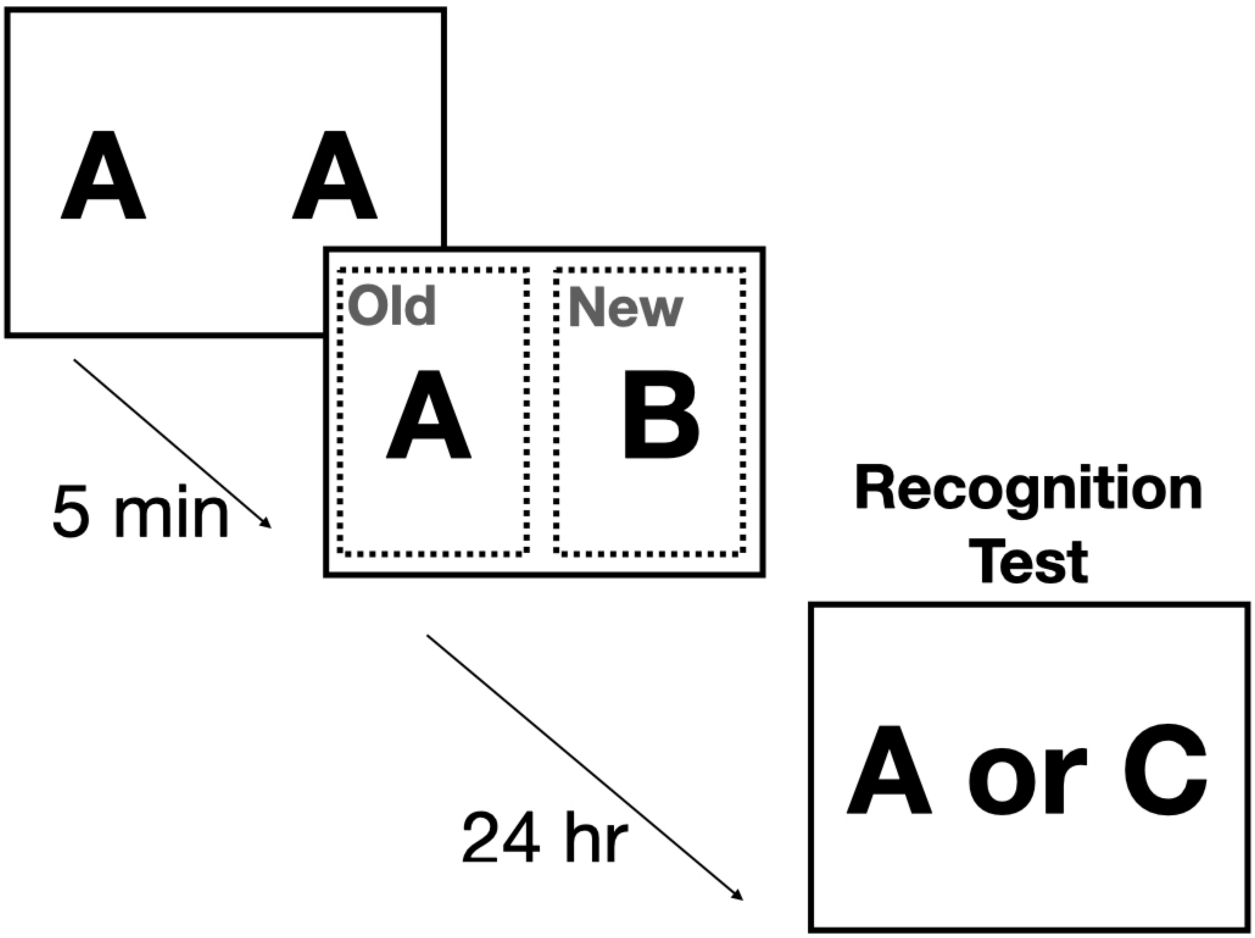


Figure 13. Schematic of the design of Experiment 3

Results

Day 1. During day 1 of the experiment, participants viewed pairs of images, one of which was always familiar and the other unfamiliar. We calculated a looking score for each participant, defined as the proportion of gaze samples in the ROI of the unfamiliar image out of all the gaze samples that were in either ROI. Gaze samples that were not in either ROI were not included in this analysis. A looking score of 0.5 indicates that participants looked equally often at the familiar and unfamiliar images, while a looking score above 0.5 indicates a preference for the unfamiliar object and a looking score below 0.5 indicate a preference for the familiar object.

Of the 1248 trials in the experiment, 78 had no fixations in either ROI, and so the looking score was unknown. We removed these trials from this analysis.

The mean looking score was 0.55 ($SD = 0.10$). This was significantly greater than 0.5 ($t(49) = 3.29$, $p = 0.00$), indicating that participants did show a preference for looking at the novel objects.

Day 2. In all of these analyses, we excluded the 16 (out of 2304) trials where the response time for the recognition judgment was greater than 10 seconds.

Participants correctly identified whether the image was familiar or unfamiliar 87.09% ($SD = 10.49$) of the time. After excluding the 3 participants who responded correctly to all images, the average confidence rating for correct responses (M = 3.51; SD = 0.41) was significantly higher than their average confidence ratings for incorrect responses (M = 2.55; SD = 0.75), t(44) = 9.36, p = 0.00 . Among the same subset of participants, response times for correct responses (M = 1,443.49, SD = 413.94) were also significantly faster than for incorrect responses (M = 2,212.65, SD = 1,733.76), t(44) = -3.43 , p = 0.00).

To see whether preferentially looking an the unfamiliar object on day 1 was correlated with confidence and response time for correct responses on Day 2, we computed the correlation coefficient between Day 1 looking scores and Day 2 confidence/RT for each participant. Following the original analysis, we transformed these values using the Fisher p-to-z transformation. Using one-sample t-tests, we found no significant difference from 0 for the correlation between looking score and confidence ratings, t(38) = 0.46, p = 0.65 (excluding the subjects who gave the same confidence judgment for all images), nor the correlation between looking score and RT, t(46) = 0.49, p = 0.63.

Calibration. To see if calibration success is correlated with the eye tracking effects, we calculated a calibration score for each participant. The calibration score was the average proportion of samples within 200 pixels of the validation points during the final validation phase before the eye tracking is performed.

Calibration scores were not correlated with looking scores, regardless of whether scores were computed using the halves coding method (see Figure @ref(fig:E3-cal-Plot-looking-score-(halves)-by-calibration) or the ROI coding method (see Figure @ref(fig:E3-cal-Plot-looking-score-(roi)-by-calibration)).

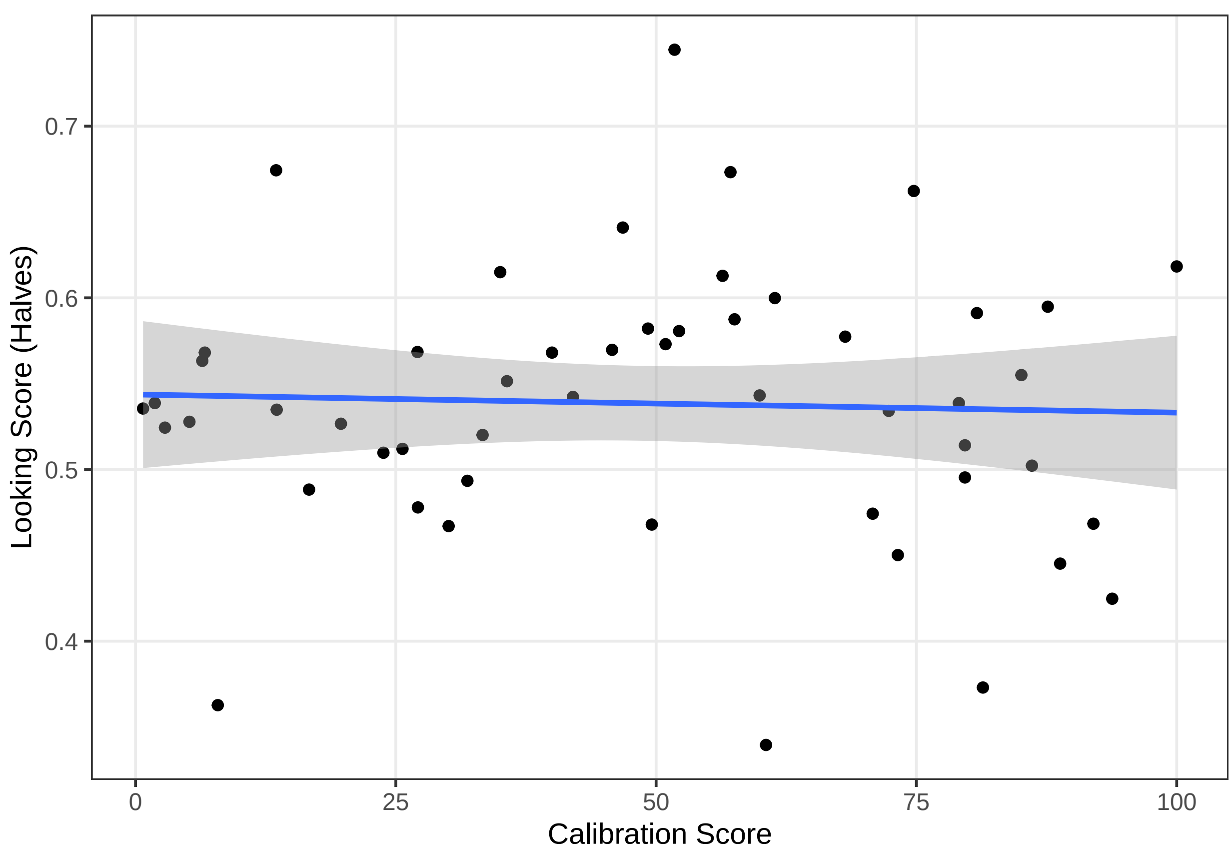


Figure 14. (#fig:E3-cal-Plot-looking-score-(halves)-by-calibration)Calibration scores plotted against looking scores using the left-vs.-right halves coding method (proportion of gaze samples to the half of the screen containing the new image out of all gaze samples).

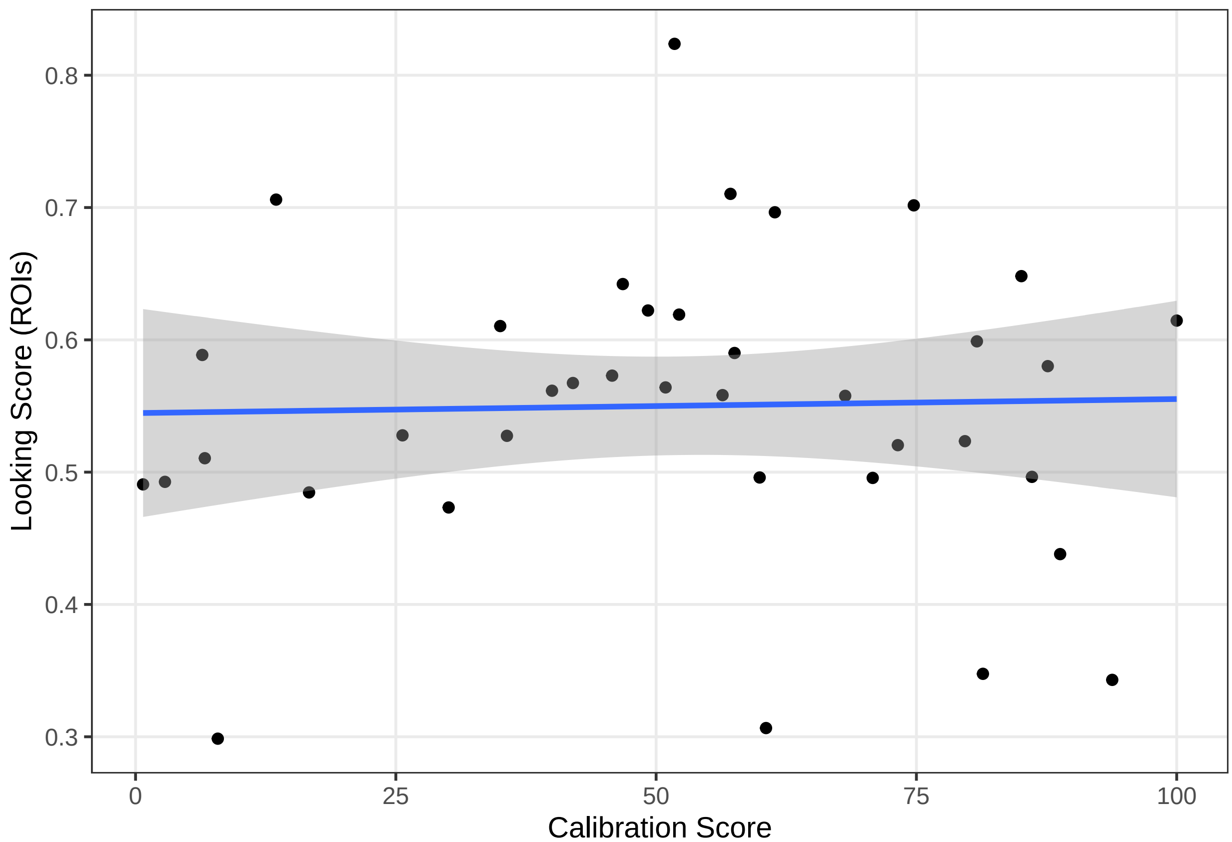


Figure 15. (#fig:E3-cal-Plot-looking-score-(roi)-by-calibration)Calibration scores plotted against looking scores using the ROI coding method (proportion of gaze samples to the new image out of all gaze samples).

We then looked at the correlation of calibration scores with the correlation between day 2 memory performance and day 1 looking scores for both kinds of behavioral and looking measures. None of the four relationships showed a significant correlation (see Figure 16).

Re-analysis After Exclusions. As is clear from the preceding figures, there was a large number of participants (N = 12) that had calibration scores under 50%. When we re-analyzed the subset that remained after those participants were excluded (N = 33), key results were aligned with the main analyses. The looking score result was upheld such that participants looked more at the new image on Day 1 (mean: 0.55 (SD = 0.10) ($t(49)$ =

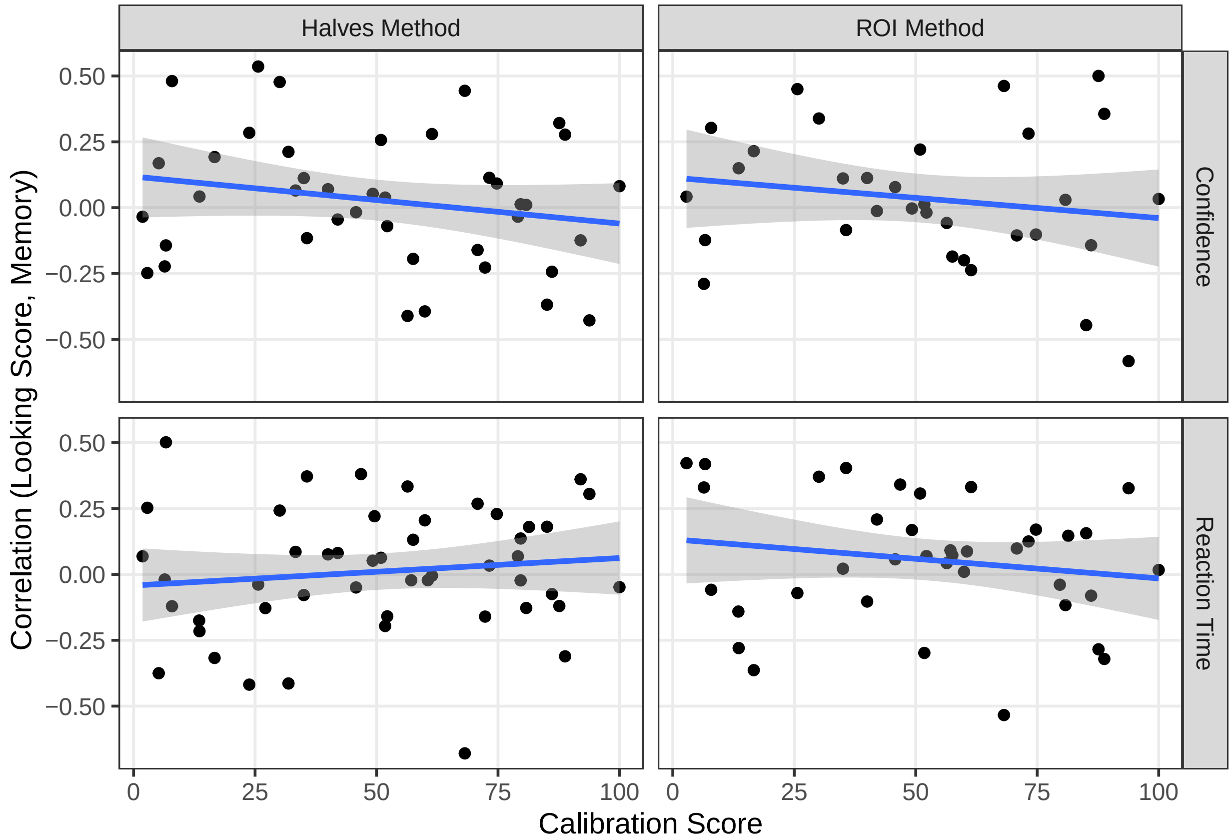


Figure 16. The relation between calibration scores and experimental effect size, defined as the correlation between Day 1 looking scores and Day 2 memory performance. Looking scores were either coded using the halves method (left panels) or ROI method (right panels); memory performance was measured using confidence ratings (top panels) or reaction time for correct recognition judgments (bottom panels).

3.29, $p = 0.00$)), but looking scores remained unrelated to Day 2 memory outcomes (confidence: t(20) = -0.32, p = 0.75; RT: t(24) = 0.10, p = 0.92).

Effects of ROIs. In the original experiment, the two objects on day 1 were presented on two separate monitors and gaze was coded by manually coding video recordings. In our replication analysis, we analyzed eye movement data using ROIs defined around the two images. In this section we explore an alternative coding of the eye movement data by coding simply left half vs. right half of the screen. The coarser coding may be more appropriate for webcam-based eyetracking.

The correlation between looking scores using the ROI method and the halves method is 0.76 (see Figure 17.

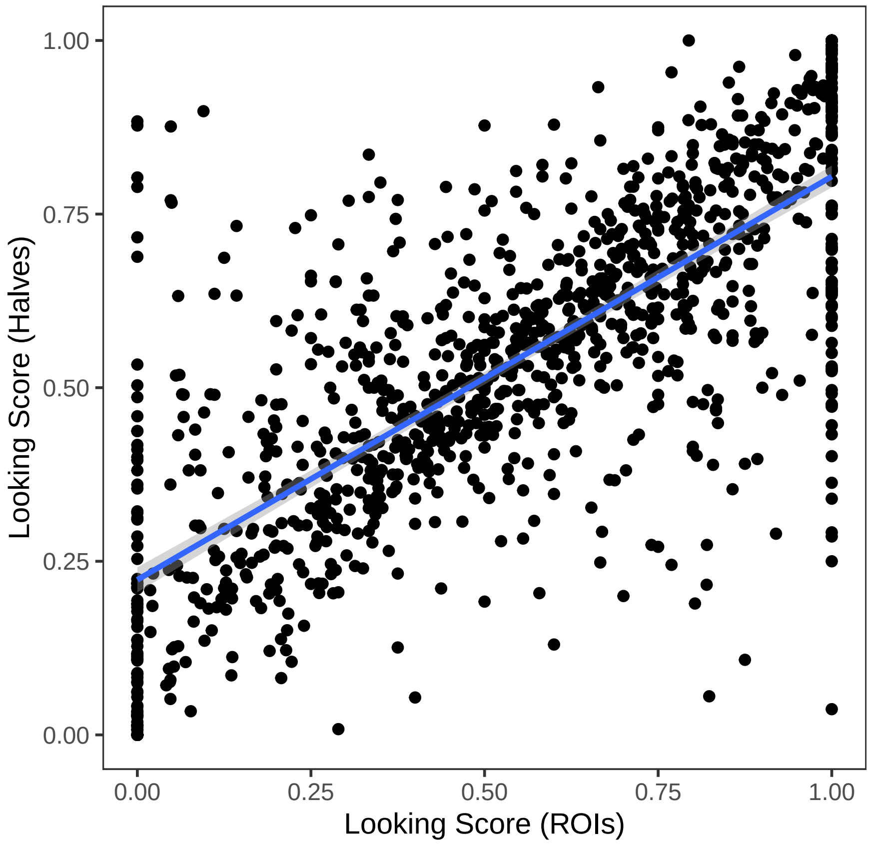


Figure 17. Correlation between looking scores calculated using ROIs and using screen halves.

Looking Scores. When looking scores are coded as left vs. right half of the screen, we find that participants looked more at the novel object. The mean looking score was 0.54 ($SD = 0.08$). This was significantly greater than 0.5, $t(50) = 3.51$, $p = 0.00$.

Correlations with Day 2 Performance. Performance on day 2 remained uncorrelated with day 1 looking scores after switching the coding of gaze. We found no significant different from 0 for the correlation between looking score and confidence ratings, t(39) = 0.74, p = 0.47 (excluding the subjects who gave the same confidence judgment for all images), nor the the correlation between looking score and RT, t(47) = 0.28, p = 0.78.

Discussion

As in Manns et al. (2000), participants looked more at novel images than previously seen images. This effect was consistent for ROIs based on the images and for the coarser ROIs based on two halves of the display. A day later, participants were also able to discriminate the images they had seen from foil images they had not seen during the previous session. However, there was no evidence that memory performance on day 2 was related to looking time on day 1. Calibration quality did not appear to impact this relationship.

Experiment 4

The fourth study was a replication attempt of Experiment 1 in Ryskin, Qi, Duff, and Brown-Schmidt (2017), which was closely modeled on Snedeker and Trueswell (2004). These studies used the visual world paradigm to show that listeners use knowledge of the co-occurrence statistics of verbs and syntactic structures to resolve ambiguity. For example, in a sentence like "Feel the frog with the feather," the phrase "with the feather" could be describing the frog, or it could be describing the instrument that should be used to do the "feeling." When both options (a frog holding a feather and a feather by itself) are available in the visual display, listeners rely on the verb's "bias" (statistical co-occurrence either in norming or corpora) to rapidly choose an action while the sentence is unfolding. .

Method

The stimuli, experimental code, and data and original analysis scripts can be found on the Open Science Framework at the following link, https://osf.io/x3c49/. The pre-registration for the study can be found at https://osf.io/3v4pg.

Participants. Fifty-seven participants were paid $2.50 for their participation. A sample size of 60 was initially chosen (but not reached in time) because we wanted to replicate the experiment with greater statistical power. Note that Ryskin et al. (2017) Experiment 1 had a sample size of 24.

Materials and Design. The images and audios presented to the participants were the same stimuli used in the original study. The critical trials were divided into modifier-biased, instrument-biased, and equibiased conditions, and the filler trials did not contain ambiguous instructions. Two lists of critical trials were made with different verb and instrument combinations (e.g., “rub” could be paired with “panda” and “crayon” in one list and “panda” and “violin” in the second list). Within each list, the same verb was presented twice but each time with a different target instrument and animal. The lists were randomly assigned to the participants to make sure the effects were not caused by the properties of the animal or instrument images used. There were 54 critical trials per list (3 verb conditions x 9 verbs per condition x 2 presentations) and 24 filler trials. The list of the 27 verbs used, along with their verb bias norms, can be found in Appendix A of the original study.

Procedure. After the eye-tracking calibration and validation (Figure 18), participants went through an audio test so they could adjust the audio on their computer to a comfortable level. Before beginning the experiment, they were given instructions that four objects would appear, an audio prompt would play, and they should do their best to use their mouse to act out the instructions. They then went through three practice trials which were followed by 54 critical trials and 24 filler trials presented in a random order.

During a trial, four pictures were displayed (target animal, target instrument, distractor animal, distractor instrument), one in each corner of the screen, and participants heard an audio prompt that contained instructions about the action they needed to act out (e.g., “Rub the butterfly with the crayon”; see Figure 19)[4]. Using their cursor, participants

[4] In the original study, the pictures appeared one by one on the screen and their names were played as they appeared. We removed this introductory portion of the trial to save time

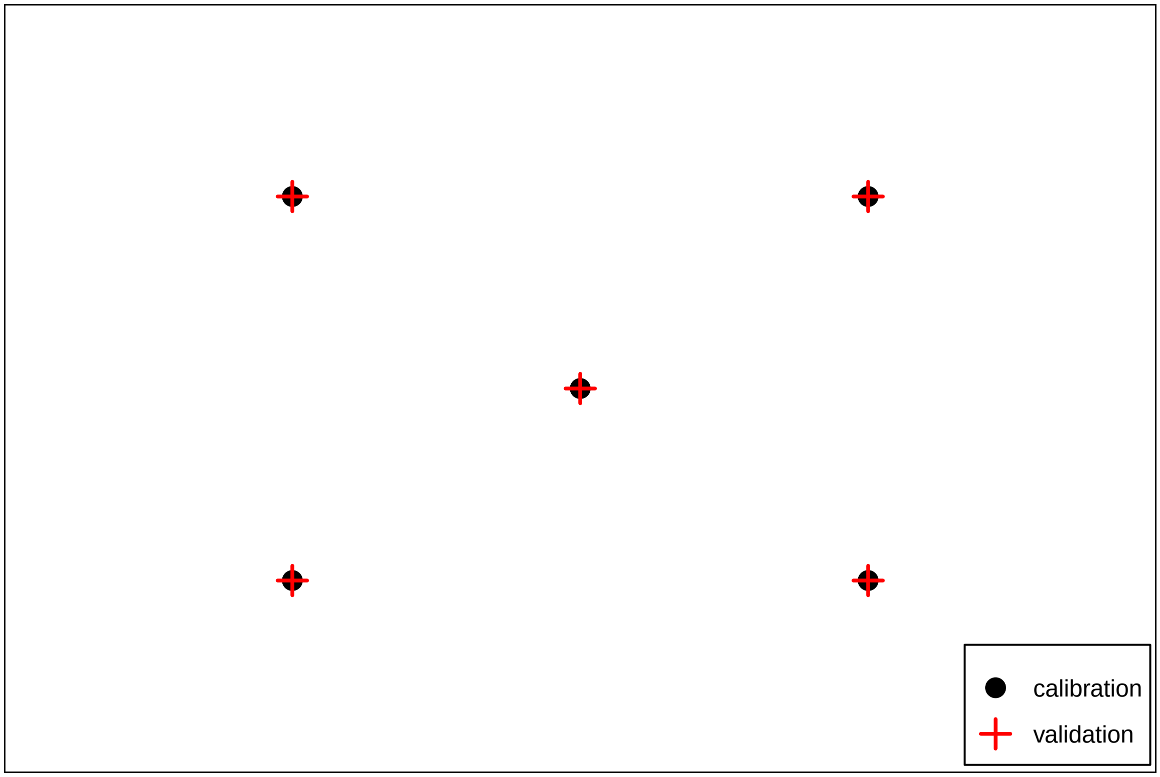


Figure 18. Calibration and validation point locations for Experiment 4. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.(In this experiment all the same locations were used for both calibration and validation.)

could act out the instructions by clicking on objects and moving them or motioning over the objects[5]. After the action was completed, the participants were instructed to press the space bar which led to a screen that said “Click Here” in the middle in order to remove bias in the eye and mouse movements from the previous trial. The experiment only allowed the participants to move on to the next trial once the audio was completely done playing and the mouse had been moved over at least one object.

[5] As opposed to the original study we recorded mouse movement instead of clicking behavior since not all of the audio prompts required clicking. For example, the sentence “locate the camel with the straw” may not involve any clicking but rather only mousing over the camel.

Figure 19. An example of a critical trial from Experiment 4 for the sentence "Rub the butterfly with the crayon." The butterfly is the target animal, the panda is the distractor animal, the crayon is the target instrument, and the violin is the distractor instrument.

Results

The location of initial mouse movements was used to assess whether the final interpretation of ambiguous sentences was biased by the verb. Figure 20 suggests that listeners were more likely to move their mouse first over the target instrument when the verb was equi-biased than when the verb was modifier-biased and even more so when the verb was instrument-biased. The opposite graded pattern can be observed for mouse movements over the target animal.

A mixed-effects logistic regression model was used to predict whether the first movement was on the target instrument with the verb bias condition as an orthogonally contrast-coded (instrument vs. equi & modifier: inst = -2/3, equi = 1/3, mod = 1/3; equi vs. modifier: inst = 0, equi = -1/2, mod = 1/2) fixed effect. Participants and items were

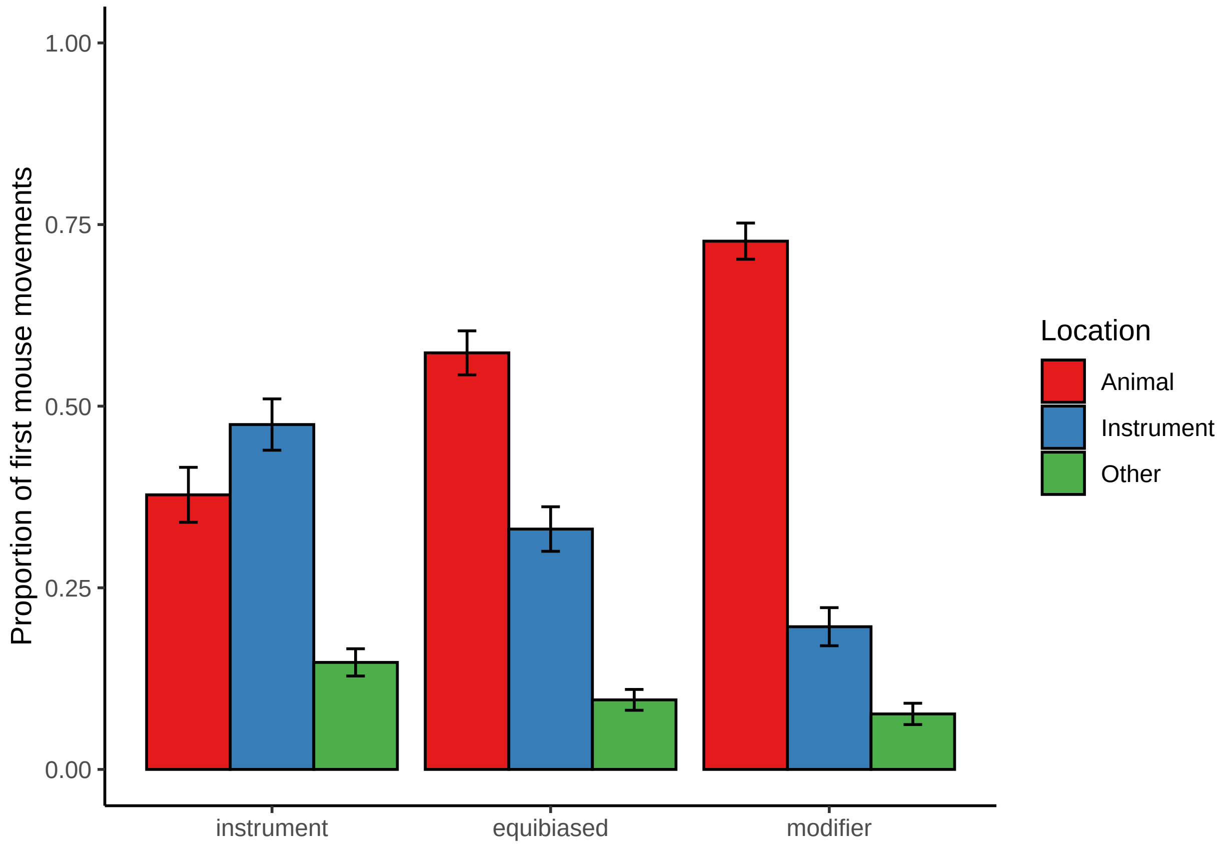


Figure 20. Proportion of first mouse movements by location and verb bias.

entered as varying intercepts with by-participant varying slopes for verb bias condition[6]. Participants were more likely to first move their mouse over target instruments in the instrument-biased condition relative to the equi-biased and modifier-biased condition (b = -1.50, SE = 0.25, $p < 0.01$). Further, participants were more likely to first move their mouse over target instruments in the equi-biased condition relative to the modifier-biased condition (b = -1.10, SE = 0.29, $p < 0.01$)

Gaze fixations were time-locked to the auditory stimulus on a trial by trial basis and categorized as being directed towards one of the four items in the display if the x, y coordinates fell within a rectangle containing the image. Figure 21 suggests that the

[6] `lme4` syntax: `glmer(is.mouse.over.instrument ~ verb_bias + (1 + verb_bias | participant) + (1 | item), family="binomial", data=d)`

participants made more fixations to the target animal when the verb was modifier-biased compared to when the the verb was equi-biased and they looked at the target animal least when the verb was instrument-biased. The pattern was reversed for looks to the target instrument.

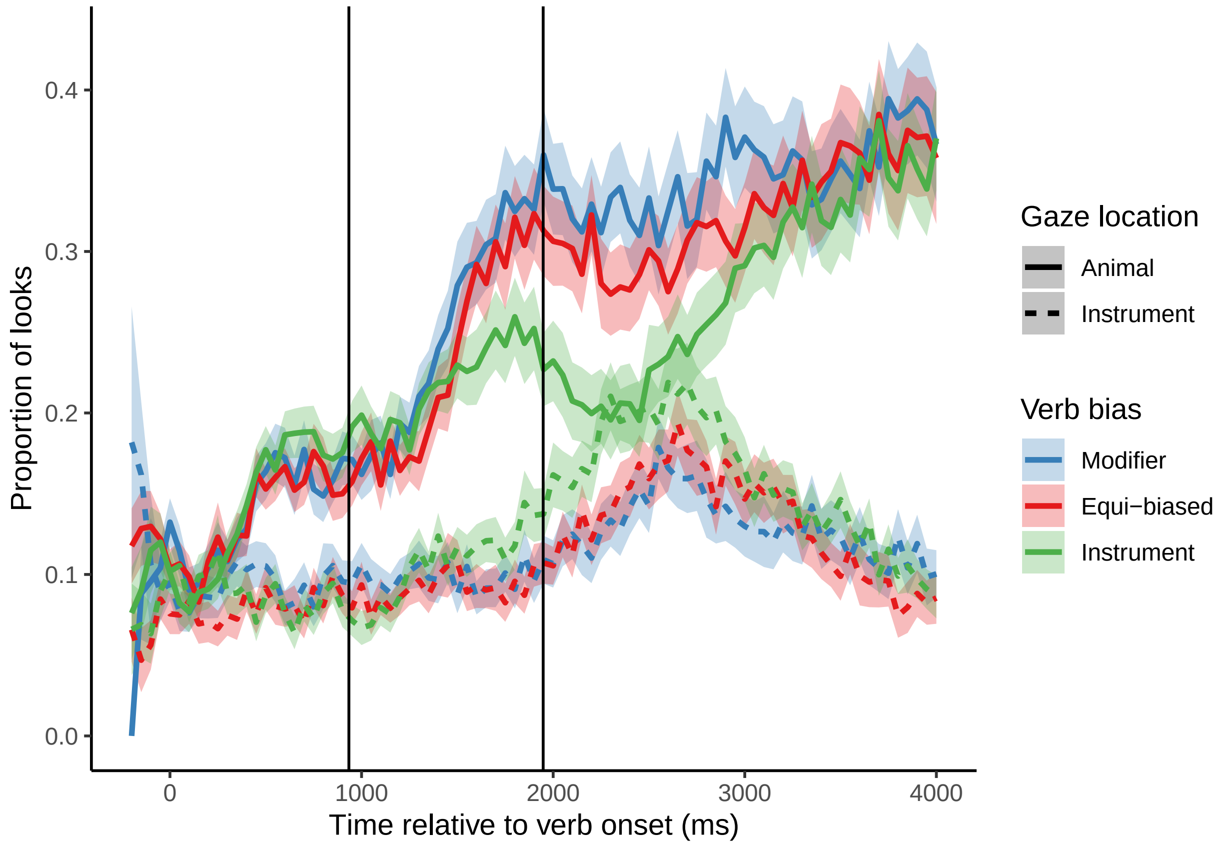


Figure 21. Timecourse of eye-gaze to target animal and target instrument by verb bias condition. Vertical lines indicate average onsets of animal and instrument offset by 200ms.

In order to assess how verb bias impacted sentence disambiguation as the sentence unfolded, the proportion of fixations was computed in three time windows: the verb-to-animal window (from verb onset + 200 ms to animal onset + 200 ms), the animal-to-instrument window (from animal onset + 200 ms to instrument onset + 200 ms), and the post-instrument window (from instrument onset + 200 ms to instrument onset + 1500ms + 200 ms). Mixed-effects linear regression models were used to predict the proportions of fixations to the target animal within each time window with the verb bias condition as an orthogonally contrast-coded (instrument vs. equi & modifier: inst = -2/3, equi = 1/3, mod = 1/3; equi vs. modifier: inst = 0, equi = -1/2, mod = 1/2) fixed effect. Participants and items were entered as varying intercepts[7]. In the *verb-to-noun* window, participants did not look more at the target animal in any of the verb bias conditions (Instrument vs. Equi and Modifier: b = -0.01, SE = 0.02, p = 0.59; Equi vs. Modifier: b = 0, SE = 0.02, p = 1). In the *noun-to-instrument* window, participants looked more at the target animal in the modifier-biased condition and equi-biased conditions relative to the instrument-biased condition (b = 0.03, SE = 0.01, $p < 0.01$) and in the modifier biased relative to the equi-biased condition (b = 0.02, SE = 0.01, $p < 0.05$). In the *post-instrument* window, participants looked more at the target animal in the modifier-biased condition and the equi-biased conditions relative to the instrument-biased condition (b = 0.08, SE = 0.02, $p < 0.01$) but not significantly so in the modifier biased condition relative to the equi-biased condition (b = 0.03, SE = 0.02, p = 0.15).

The web version of the study qualitatively replicates the action and eye-tracking results of the original dataset (Ryskin et al., 2017). The mouse click results from both studies are summarized in Figure 22. The quantitative patterns of clicks were similar to those observed in the original dataset, though for Instrument-biased verbs, clicks were closer to evenly split between the animal and the instrument relative to the in-lab study where they were very clearly biased toward the instrument.

The eye-tracking results from both studies are summarized in Figure 23. For simplicity, and to reflect the dependent variable used in analyses, we average the proportion of fixations to the target animal within each time window. Though the qualitative patterns are replicated, proportions of fixations to the target animal were much

[7] `lme4` syntax: `lmer(prop.fix.target.animal ~ verb_bias + (1 + verb_bias | participant) + (1 | item), data=d)`. A model with by-participant varying slopes for verb bias condition was first attempted but did not converge.

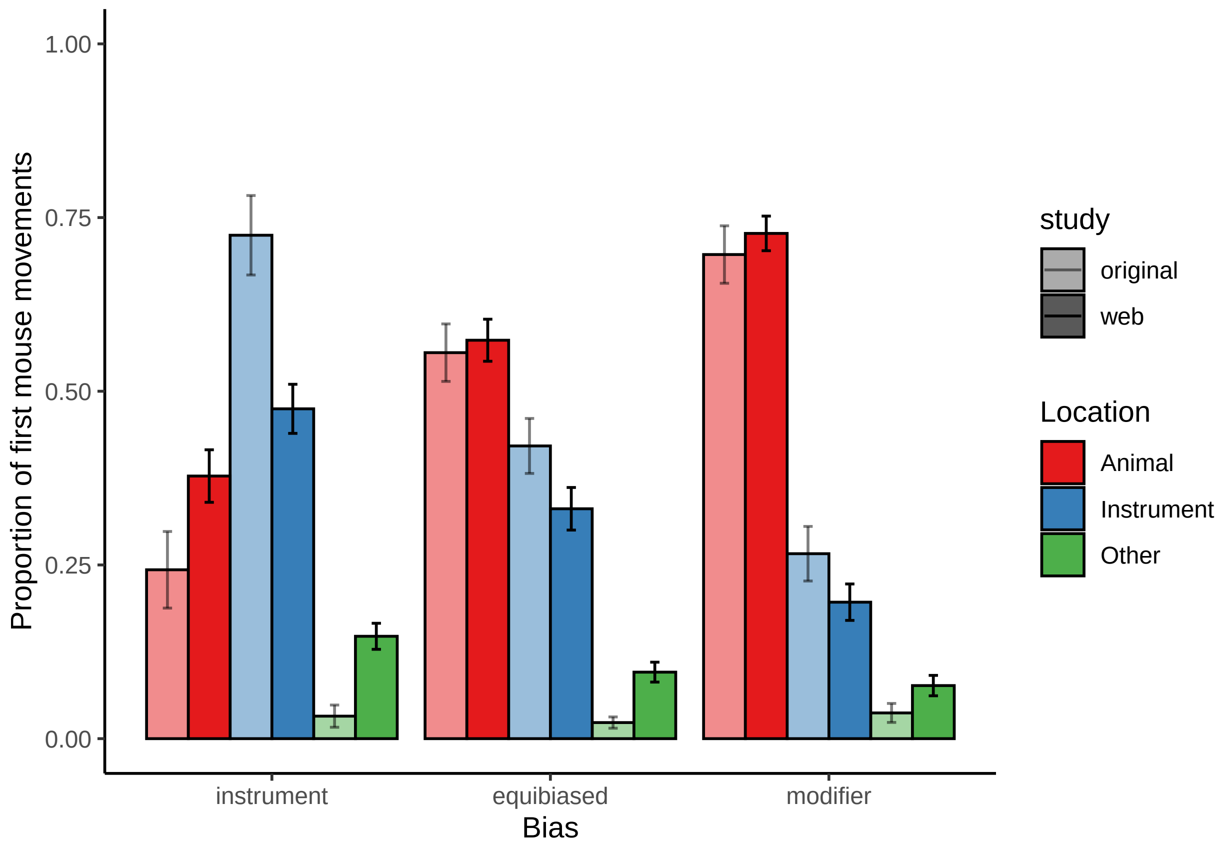


Figure 22. Proportion of first mouse movements by location and verb bias in the original dataset (Ryskin et al., 2017) and the current data collected online.

lower in the web version of the study. This may reflect the fact that participants in the web study are less attentive and/or the quality of the webgazer eye-tracking system is lower, relative to the Eyelink 1000 which was used for the original study.

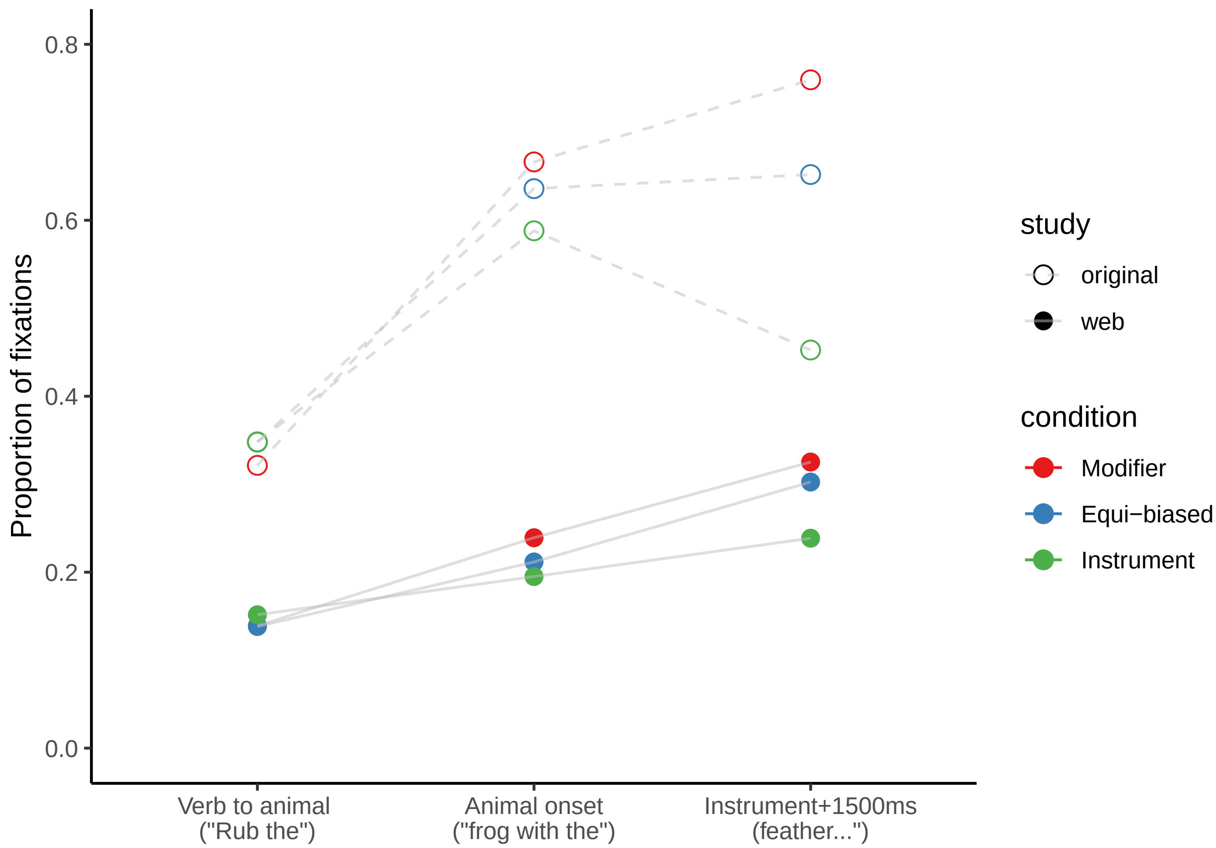
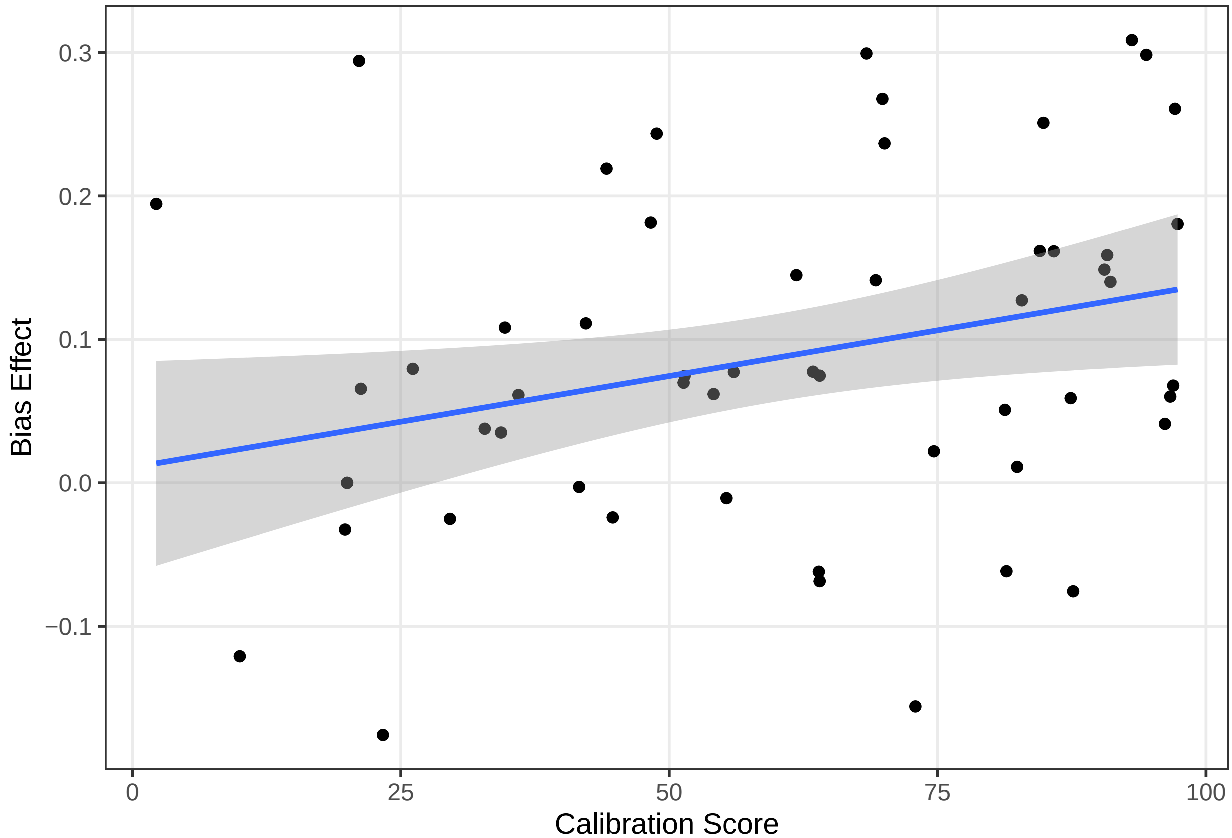


Figure 23. Proportion of target fixations by verb bias in the original dataset (Ryskin et al., 2017) and the current data collected online. Error bars reflect bootstrapped 95% CIs over subject means

Calibration.



Participants' calibration quality, measured as the mean percentage of fixations that landed within 200 pixels of the calibration point, varied substantially (between 2.22 and 97.36%). The quality of a participant's calibration significantly correlated with the participant's effect size (*Pearson's* r= 0.29, $p < 0.05$). The difference in target animal fixation proportions between modifier and instrument conditions was higher for participants with better calibration (see Figure **??**).

Re-analysis After Exclusions. Replicating the linear mixed-effects analysis (in the post-instrument onset time window only) on a subset of 35 participants with calibration quality >50% suggests that the effect of verb bias condition was larger in this subset than in the full dataset. Participants looked more at the target animal in the modifier-biased condition and the equi-biased conditions relative to the instrument-biased condition ($b = 0.10$, $SE = 0.02$, $p < 0.001$) but not significantly so in the modifier biased condition relative to the equi-biased condition ($b = 0.02$, $SE = 0.02$, $p = 0.29$).

Effects of ROIs. Eye-tracking on the web differs critically from in-lab eye-tracking in that the size of the display differs across participants. Thus the size of the ROIs differs across participants. The current version of the web experiment used a bounding box around each image to determine the ROI. This approach is flexible and accomodates variability in image size, but may exclude looks that are directed at the image but fall outside of the image (due to participant or eye-tracker noise) as shown in Figure 24a. Alternatively, The display can be split into 4 quadrants which jointly cover the entire screen (see Figure 24b).

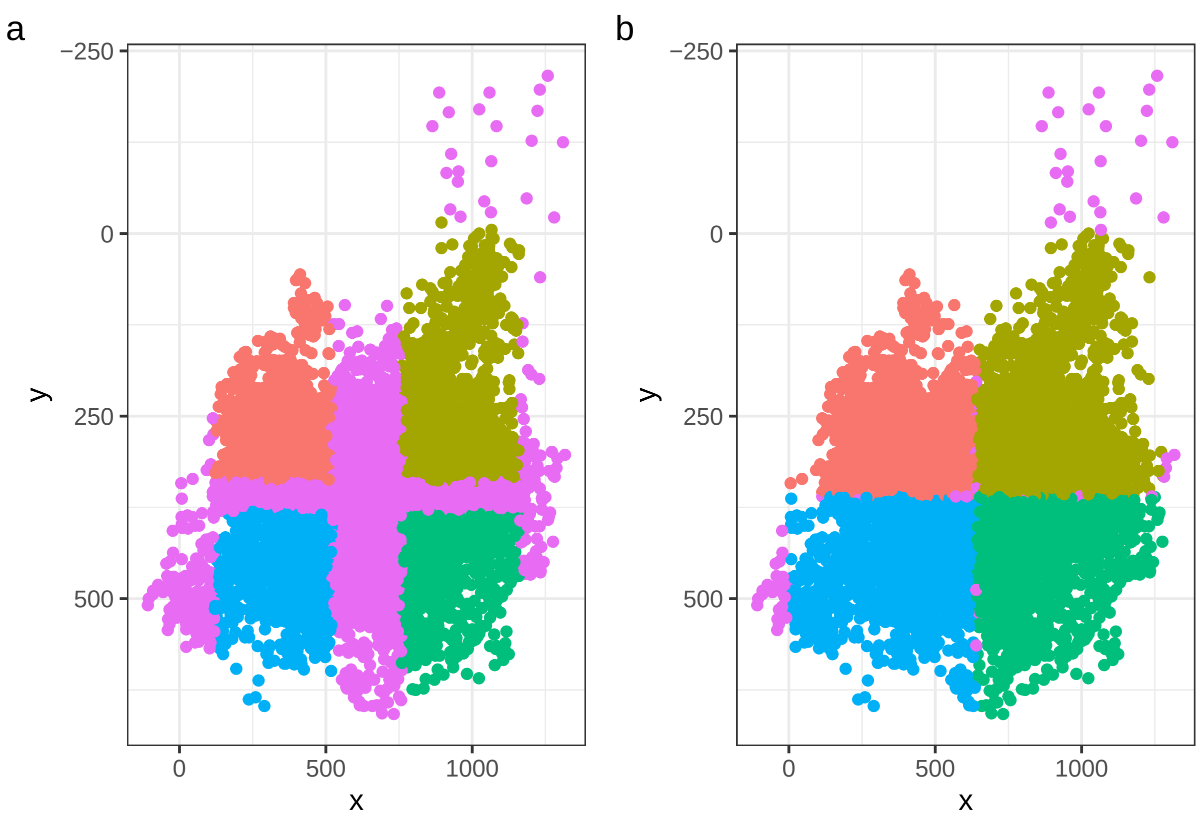


Figure 24. Example participant's gaze coordinates categorized into ROIs based on a) image bounding boxes and b) screen quadrants. Magenta points indicate looks that were not categorized into an ROI.

Categorizing gaze location based on which of the four quadrants of the screen the

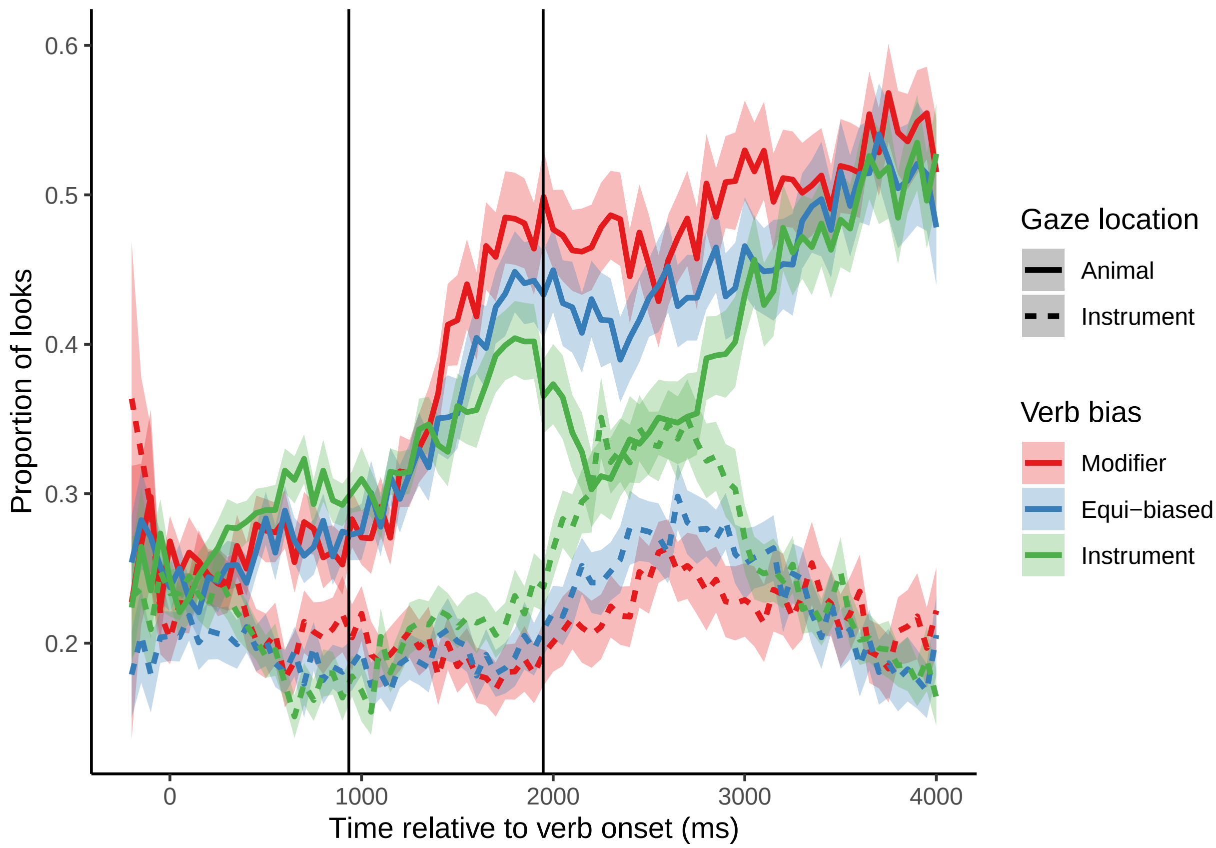


Figure 25. Timecourse of eye-gaze to target animal and target instrument by verb bias condition with gaze categorized based on which quadrant of the screen the coordinates fall in (as opposed to a bounding box around the image). Vertical lines indicate average onsets of animal and instrument offset by 200ms.

coordinates fell in, increases the overall proportions of fixations (see Figure 25). In the *post-instrument* window, participants looked more at the target animal in the modifier-biased condition and the equi-biased conditions relative to the instrument-biased condition ($b = 0.08$, $SE = 0.02$, $p < 0.01$) and marginally so in the modifier biased condition relative to the equi-biased condition ($b = 0.04$, $SE = 0.02$, $p = 0.05$). Effect size estimates appeared somewhat larger and noise was somewhat reduced when using the quadrant categorization relative to the bounding box-based ROIs.

Discussion

As in Ryskin et al. (2017) and Snedeker and Trueswell (2004), listeners' gaze patterns during sentences with globally ambiguous syntactic interpretations differed depending on the bias of the verb (i.e., modifier-, instrument-, or equi-). For modifier-biased verbs, participants looked more quickly at the target animal and less at the potential instrument than for instrument-biased verbs (and equi-biased verbs elicited a gaze pattern between these extremes). This pattern was stronger for those who achieved higher calibration accuracy and when quadrant-based ROIs were used compared to image-based ROIs.

Experiment 5

The fifth study was a replication attempt of Shimojo, Simion, Shimojo, and Scheier (2003), which found that human gaze is actively involved in preference formation. Separate sets of participants were shown pairs of human faces and asked either to choose which one they found more attractive or which they felt was rounder. Prior to making their explicit selection, participants were increasingly likely to be fixating the face they ultimately chose, though this effect was significantly weaker for roundness discrimination.

Note that Shimojo and colleagues compare five conditions, of which we replicate only the two that figure most prominently in their conclusions: the "face-attractiveness-difficult task" and the "face-roundness task".

Method

All stimuli, experiment scripts, data, and analysis scripts are available on the Open Science Framework at https://osf.io/eubsc/. The study pre-registration is available at https://osf.io/tv57s.

Participants. Fifty participants for the main task were recruited on Prolific and were paid $10/hour. For one subject (roundess task group), eye gaze data failed to record.

We ended up with 25 participants in the attractiveness condition and 24 in the roundness condition. The original sample size in Shimojo et al. (2003) was 10 participants total.

Materials and Design. The faces in our replication were selected from a set of 1,000 faces within the Flickr-Faces-HQ Dataset (the face images used in Shimojo et al. were from the Ekman face database and the AR face database). These images were chosen because the person in each image was looking at the camera with a fairly neutral facial expression and appeared to be over the age of 18. Twenty-seven participants were recruited on Prolific to participate in stimulus norming (for attractiveness). They each viewed all 172 faces and were asked to rate them on a scale from 1 (less attractive) to 7 (more attractive) using a slider. Faces were presented one at a time and in a random order for each participant. Data from three participants were excluded because their modal response made up more than 50% of their total responses, for a total of 24 participants in the norming.

Following Shimojo et al., 19 face pairs were selected by identifying two faces that 1) had a difference in mean attractiveness ratings that was 0.25 points or lower and 2) matched in gender, race, and age group (young adult, adult, or older adult).

Participants were presented with each of the 19 face pairs; task condition was manipulated between subjects, such that half of participants made judgments about facial attractiveness and the other half made judgments about facial shape.

Procedure. At the beginning of the experimental task, participants completed a 9-point eye-tracker calibration (each point appeared 3 times in random order) and 3-point validation. The validation point appeared once at center, middle left, and middle right locations in random order (see Figure 26).

During each trial of the main task, two faces were displayed on the two halves of the screen, one on the left and one on the right (as in Figure 27). In the attractiveness task, participants were asked to chose the more attractice face in the pair and in the shape

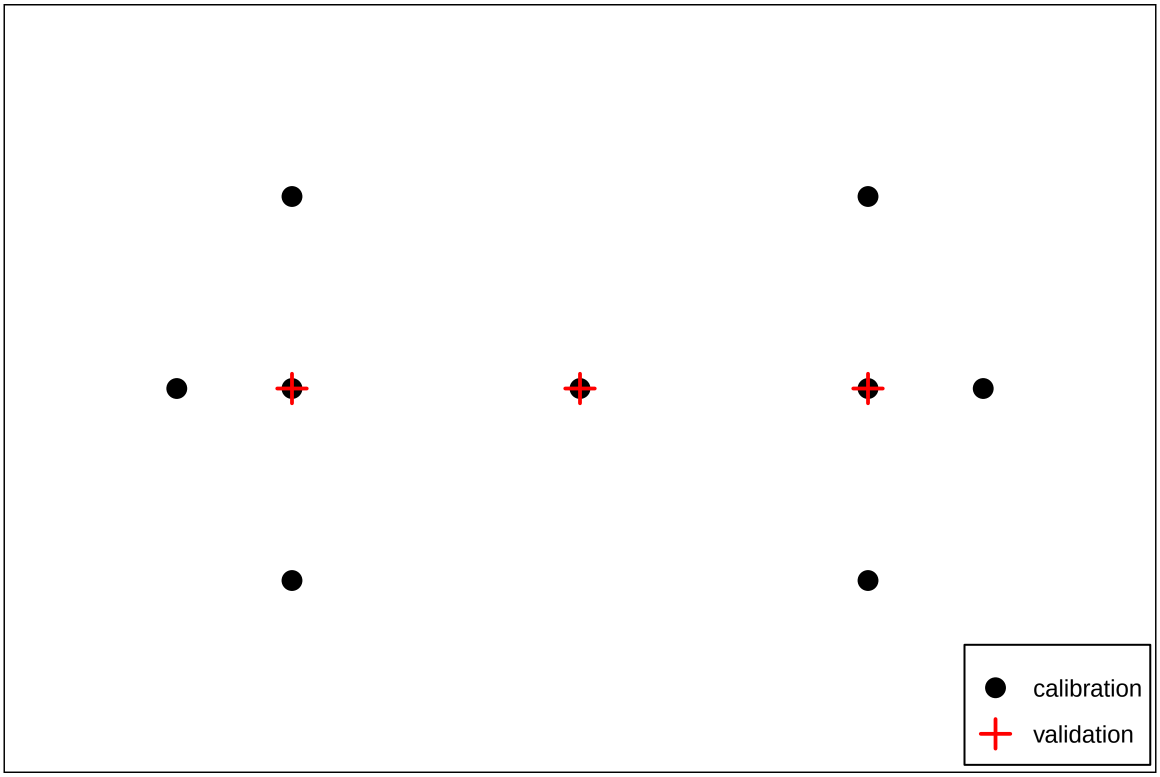


Figure 26. Calibration and validation point locations for Experiment 5. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.

judgment task participants were asked to pick the face that appeared rounder. They pressed the "a" key on their keyboard to select the face on the left and the "d" key to select the face on the right. A fixation cross appeared in the center of the screen between each set of faces. Participants were asked to look at this fixation cross in order to reset their gaze in between trials. The order of the 19 face pairs was random for each participant.

Results

In the original study, a video-based eye tracker was used. The eye movements of participants were recorded with a digital camera downsampled to 33.3 Hz, with eye position determined automatically with MediaAnalyzer software. In our study, subjects supplied their own cameras, so hardware sampling rate varied. However, data was collected at 20 Hz.

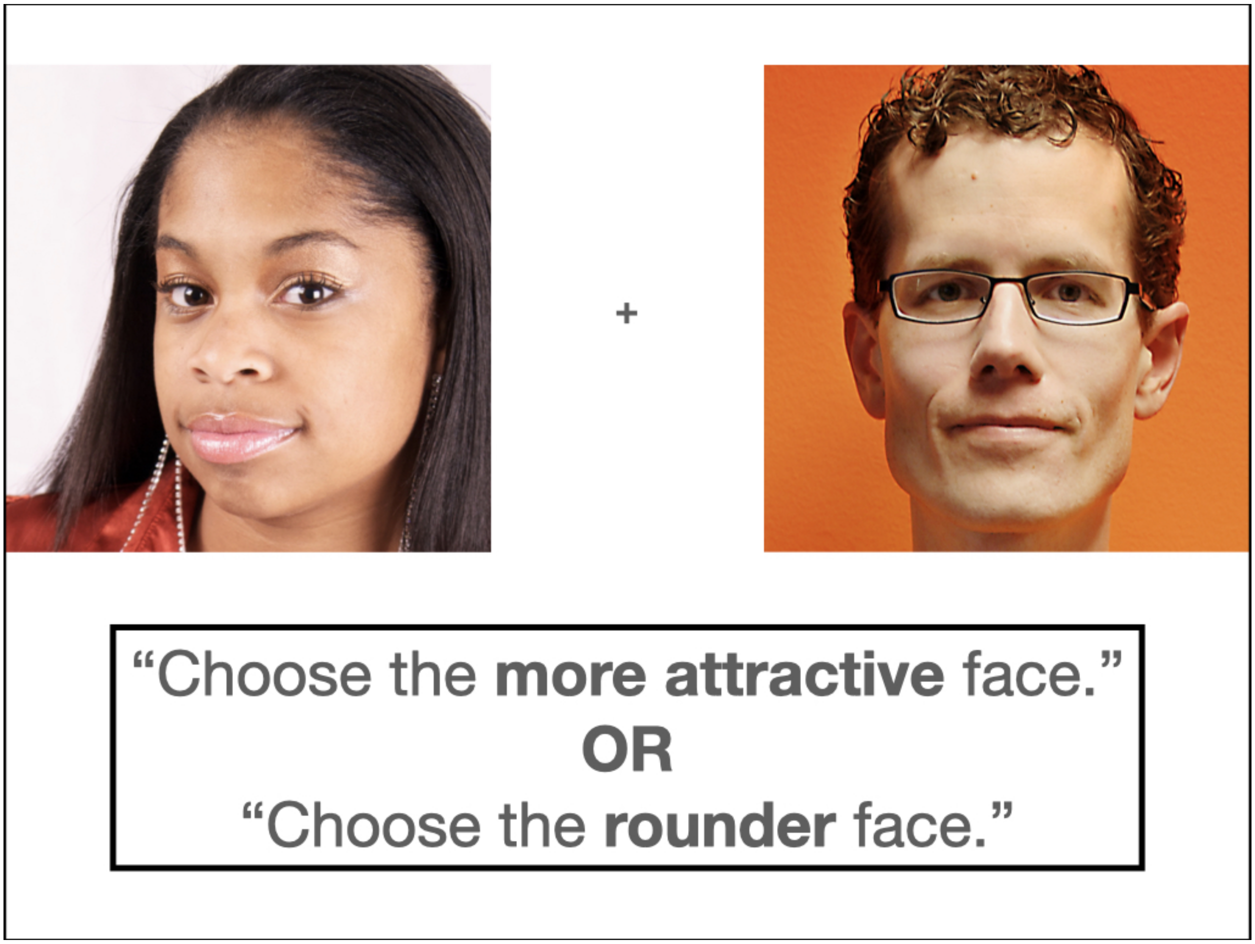


Figure 27. An example of a critical trial from Experiment 5 (text did not appear on each screen).

Due to large variation in response time latency, Shimojo and colleagues analyzed eye gaze for the 1.67 seconds prior to the response. This duration was one standard deviation of the mean response time, ensuring that all timepoints analyzed have data from at least 67% of trials. In our dataset, one standard deviation amounts to 1.95 seconds. We then binned eyegaze data into 50 ms bins rather than the 30 ms bins used by Shimojo and colleagues, reflecting the different sampling rates.

Following Shimojo and colleagues, data for each condition were fit using a four-parameter sigmoid (Fig. 28). These fit less well than in the original paper for both the attractiveness judgment ($R^2 = 0.80$ vs. 0.91) and the roundness judgment ($R^2 = 0.36$

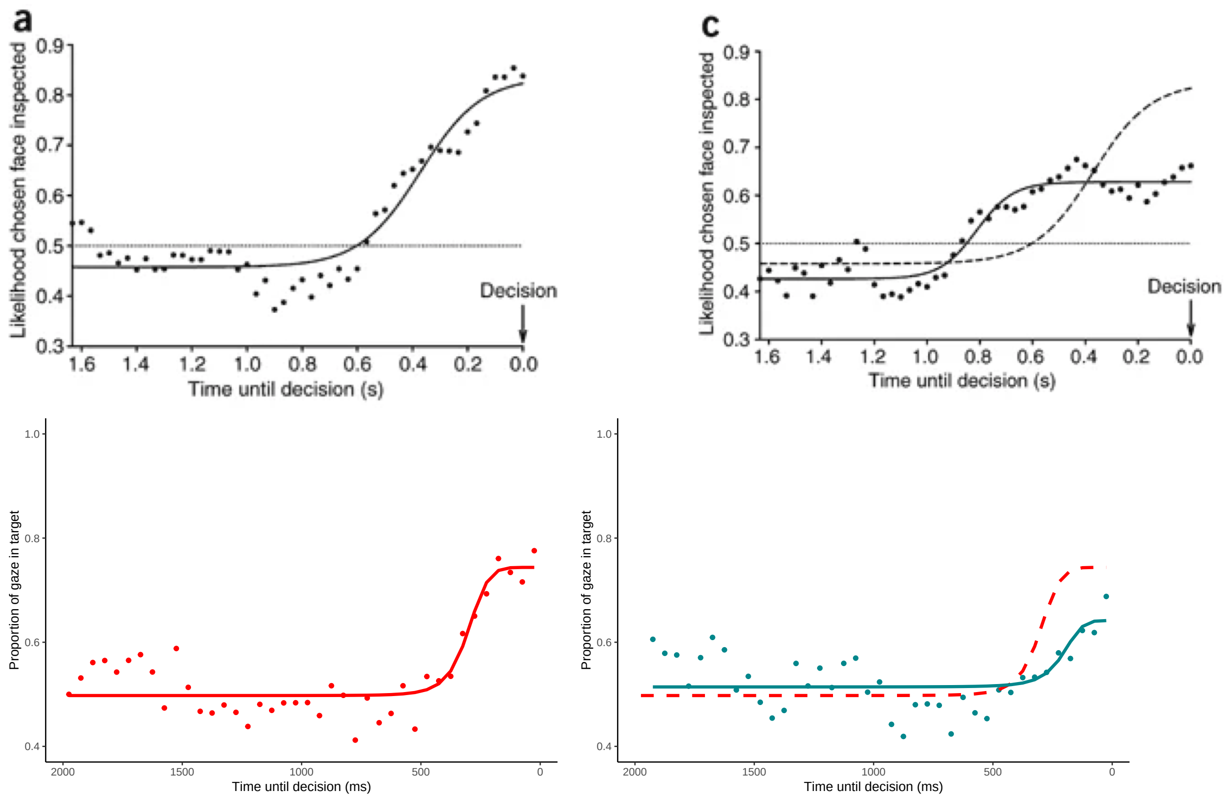


Figure 28. Primary results from Exp. 5. *Top* shows the original results from Shimojo and colleagues (Figures reprinted with permission[TODO]). The attractiveness judgment along with the best-fitting sigmoid is shown in the *top left*. Results for the roundness judgment are show in the *top right*, with the best-fitting sigmoid for the attractiveness judgment depicted in a dashed line for comparison (*top right*). (*Bottom*) shows the analogous results from the replication, with the attractiveness judgments on the *bottom left* and the roundness judgments on the *bottom right*. Again, the best-fitting sigmoid for the attractiveness judgments are plotted with a dashed line alongside the roundness results, for purposes of comparison.

vs. 0.91).

From these curves, Shimojo and colleagues focus on two qualitative findings. First, they note a higher asymptote for the attractiveness discrimination task relative to roundness discrimination. Qualitatively, this appears to replicate. However, their statistical analysis – a Kolmogorov-Smirnov test for distance between two distributions – is not significant (D = 0.17, p = 0.55), though it should be noted that this is a very indirect statistical test of the hypothesis and probably not very sensitive.

The second qualitative finding they note is that the curve for the roundness judgment "saturates" (asymptotes) earlier than the curve for the attractiveness judgment. They do not present any statistical analyses, but it is clear qualitatively that the result does not replicate.

Calibration. As in the previous experiments, calibration score was defined as the average proportion of samples within 200 pixels of the validation point during the final validation phase before the eye tracking is performed. Where participants required more than one calibration (N=14), only the final calibration was considered.

To determine whether calibration accuracy influenced our key effects, we calculated the proportion of samples during the task in which the participant was fixating the face they ultimately chose. Calibration accuracy significantly correlated with fixations in both the attractiveness condition (r = 0.50 [0.12, 0.75], p = 0.01) and the roundness condition (r = 0.25 [-0.18, 0.60], p = 0.25). Inspection of Fig. 29 reveals that this correlation is due to a handful of participants with calibration values below 50%.

Re-analysis After Exclusions. As in the previous experiments, we re-analyzed the data, removing the participants whose calibration accuracy was not greater than 50%. This slightly improved the fits of the sigmoids (Attractiveness: R^2 = 0.78; Roundness: R^2 = 0.59). However, the difference between sigmoids remained non-significant using the Kolmogorov-Smirnov test (D = 0.20, p = 0.33). Descriptively, the results do not look

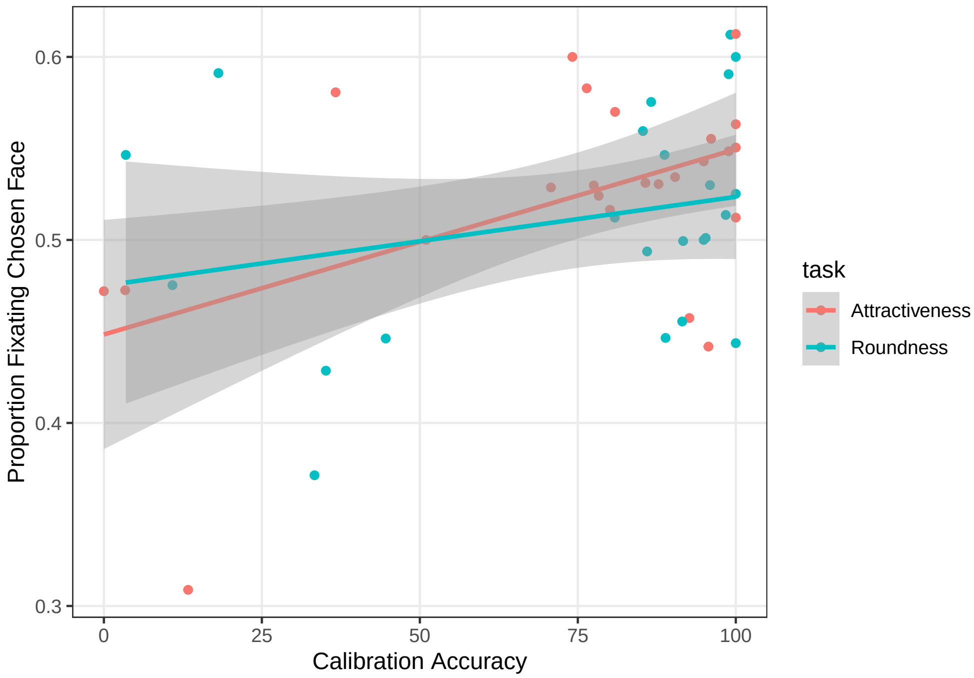


Figure 29. Correlation between calibration accuracy (x-axis) and proportion of samples fixating target (y-axis) in Exp. 5.

substantially different (Fig. 30).

Effects of ROIs. In the original experiment, eye gazes that did not directly fixate one or other of the faces were excluded. In this section we explore an alternative coding of the eye movement data by coding simply left half vs. right half of the screen. The coarser coding may be more appropriate for webcam-based eyetracking.

Only a small percentage of samples (7.00%) involved looks to anything other than one of the two faces. Thus, not surprisingly, the correlation between percentage of time spent fixating the to-be-chosen face using the ROI method and the halves method was near ceiling ($r = 0.97$ [0.97, 0.98], $p = 0$). Since the choice of method had almost no effect on whether participants were coded as fixating one face or the other, we did not further investigate the effect of method choice on the analytic results.

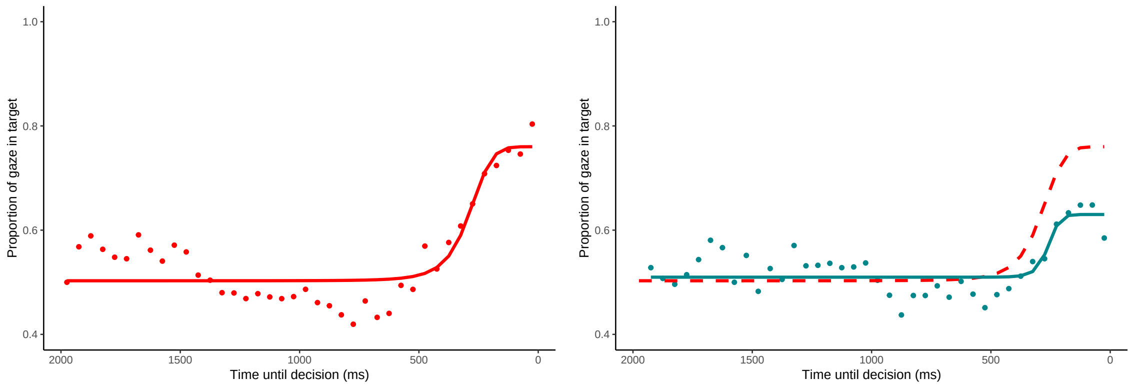


Figure 30. Revised results for Exp. 5 after removing low-calibration accuracy participants. *Left*: Eyegaze during attractiveness judgments, along with the best-fitting sigmoid. *Right*: Eyegze during roundness judgments, along with best-fitting sigmoid (best-fitting sigmoid for attractiveness is re-plotted with a dashed line for comparison).

Discussion

Qualitatively, the results are similar to those of Shimojo et al., such that participants look more at the option that they ultimately choose. This gaze bias appears to be stronger for decisions about face attractiveness than shape, though this is not supported by the statistical analysis approach used in the original paper. The gaze patterns remained consistent for participants with better calibration accuracy.

General Discussion

We conducted five attempted replication studies using different experimental paradigms from across the cognitive sciences. All were successfully implemented in `jsPsych` using the `webgazer` plugin, but replication success was mixed. Experiment 1 had the smallest ROIs due to the use of an integrated visual scene with five to six ROIs of varying size per scene, as opposed to ROIs corresponding to display halves or quadrants. Both attempts to replicate Altmann and Kamide (1999) were unsuccessful, despite the success of previous in-lab replications using infrared eye-tracking (e.g. James et al., 2023). A previous conceptual replication of this paradigm using webcam-based eye-tracking (Prystauka et al., 2023) was successful but used a four-quadrant visual world paradigm, rather than the “naturalistic” scenes used in the original study and in the current replication attempts. It is worth noting that removing variability related to participant environments (by conducting the webcam-tracking study in the lab) did not appear to improve the sensitivity of the paradigm. The primary limitation is likely to be the size of the ROIs.

Experiment 2 used the four quadrants of the participant’s screen as ROIs. As in Johansson and Johansson (2014) and Spivey and Geng (2001), participants spontaneously looked to blank ROIs which previously contained to-be-remembered pictures. These results appeared to be robust to calibration quality. An additional manipulation, instructing participants to keep gaze fixed on a central point, was not successful. One possibility is that participants are less motivated to follow such instructions when an experimenter is not present in the same room with them. It may be possible to improve performance by emphasizing that this is an important aspect of the experiment or by providing additional training/practice in keeping the eyes still on one particular point.

Experiment 3 used two large ROIs (halves of the display in one analysis) and successfully replicated the novelty preference in terms of gaze duration shown in Manns et al. (2000). However, the subtler relationship between gaze duration and recognition memory on Day 2 was not replicated, despite the fact that participants were able to discriminate pictures they had seen from those they hadn’t seen during that delayed test. Calibration quality did not appear to impact this relationship. More work is needed to understand whether delay manipulations can be practically combined with webcam eye-tracking.

Experiment 4 used the four quadrants of the participant’s screen as ROIs. As in Ryskin et al. (2017), listeners used knowledge of the co-occurrence statistics of verbs and syntactic structures to resolve ambiguous linguistic input (“Rub the frog with the mitten”).

Across multiple time windows, participants looked more at potential instruments (mitten), when the verb (rub) was one that was more likely to be followed by a prepositional phrase describing an instrument with which to perform the action , as opposed to describing the recipient of the action (frog). Despite the qualitative replication of past findings, the overall rates of looks to various objects were much lower than in an in-lab study using infrared eye-tracking. This reduction may be related to measurement quality: effect sizes were greater for participants with higher calibration accuracy. Using the full quadrants as ROIs, rather than bounding boxes around the four images, also appeared to improve the measurement of the effect. Crucially, there was no evidence of a delay in the onset of effects relative to in-lab work, indicating that the modifications to `webgazer` that are made within the `jsPsych` plug-in successfully address the issues noted by Dijkgraaf, Hartsuiker, and Duyck (2017) and suggesting that this methodology can be fruitfully used to investigate research questions related to the timecourse of processing.

Experiment 5, similar to Experiment 3, used two large ROIs (or halves of the display). As in Shimojo et al. (2003) and in the recent webcam-based replication by Yang and Krajbich (2021), we saw that participants looked more at the face or shape that they ultimately chose during a judgment task. This gaze bias appears to be stronger for decisions about face attractiveness than shape, though this effect was not statistically significant.

In sum, the `webgazer` plug-in for `jsPsych` can be fruitfully used to conduct a variety of cognitive science experiments on the web, provided the limitations of the methodology are carefully considered. Studies with ROIs that take up half or a quarter of the participant's display, which encompasses a large number of common paradigms, are very likely to be successful, even when testing questions related to the timecourse of processing. However, the smaller the ROIs, the more important the calibration becomes. For instance, studies with four ROIs may want to exclude data from participants with less than 75% validation accuracy, whereas studies using two halves of the display as ROIs may not need to be so conservative. Studies with smaller ROIs (see Experiment 1) may not be appropriate for webcam eye-tracking in its current form.

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

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407.

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

Barth, M. (2022). *tinylabels: Lightweight variable labels.* Retrieved from https://cran.r-project.org/package=tinylabels

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

Bates, D., & Maechler, M. (2021). *Matrix: Sparse and dense matrix classes and methods.* Retrieved from https://CRAN.R-project.org/package=Matrix

Bolker, B., & Robinson, D. (2020). *Broom.mixed: Tidying methods for mixed models.* Retrieved from https://CRAN.R-project.org/package=broom.mixed

Burton, L., Albert, W., & Flynn, M. (2014). A comparison of the performance of webcam vs. Infrared eye tracking technology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*, 1437–1441. SAGE Publications Sage CA: Los Angeles, CA.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., … Borges, B. (2021). *Shiny: Web application framework for r.* Retrieved from https://CRAN.R-project.org/package=shiny

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

De Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a web browser? Comparing response times collected with JavaScript and psychophysics toolbox in a visual search task. *Behavior Research Methods*, *48*, 1–12.

Degen, J., Kursat, L., & Leigh, D. 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*.

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.

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the internet in reaching more diverse samples. *Behavioral and Brain Sciences*, *33*(2-3), 94.

Hartshorne, J. K., Leeuw, J. R. de, Goodman, N. D., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with pushkin. *Behavior Research Methods*, *51*, 1782–1803.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), 743–747.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). Lab. Js: A free, open, online study builder. *Behavior Research Methods*, 1–18.

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

James, A. N., Minnihan, C. J., & Watson, D. G. (2023). Language experience predicts eye movements during online auditory comprehension. *Journal of Cognition*, *6*(1).

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

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Li, W., Germine, L. T., Mehr, S. A., Srinivasan, M., & Hartshorne, J. (2024). Developmental psychologists should adopt citizen science to improve generalization and reproducibility. *Infant and Child Development*, *33*(1), e2348.

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

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., et al.others. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748.

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between JSON data and r objects. *arXiv:1403.2805 [Stat.CO]*. Retrieved from https://arxiv.org/abs/1403.2805

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

Passell, E., Strong, R. W., Rutter, L. A., Kim, H., Scheuer, L., Martini, P., ... Germine, L. (2021). Cognitive test scores vary with choice of personal digital device. *Behavior Research Methods*, *53*(6), 2544–2557.

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. *Behavior Research Methods*, 1–19.

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/

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372.

Reinecke, K., & Gajos, K. Z. (2015). LabintheWild: Conducting large-scale online experiments with uncompensated samples. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1364–1378.

Richardson, D. C., & Spivey, M. J. (2004). Eye tracking: Research areas and applications. In G. Wnek & G. Bowlin (Eds.), *Encyclopedia of biomaterials and biomedical engineering* (Vol. 572). New York: Marcel Dekker.

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

Ryskin, R., Salinas, M., Piantadosi, S., & Gibson, E. (2023). Real-time inference in communication across cultures: Evidence from a nonindustrialized society. *Journal of Experimental Psychology: General*, *152*(5), 1245.

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

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, *6*(12), 1317–1322. https://doi.org/10.1038/nn1150

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2021). *Afex: Analysis of factorial experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Skovsgaard, H., Agustin, J. S., Johansen, S. A., Hansen, J. P., & Tall, M. (2011). Evaluation of a remote webcam-based eye tracker. *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications*, 1–4.

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. *Behavior Research Methods*, 1–19.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103–130.

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

Spivey, M. J., & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research*, *65*(4), 235–241. https://doi.org/10.1007/s004260100059

Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A., et al.others. (2024). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *Infancy*, *29*(1), 31–55.

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.

Van der Cruyssen, I., Ben-Shakhar, G., Pertzov, Y., Guy, N., Cabooter, Q., Gunschera, L. J., & Verschuere, B. (2023). The validation of online webcam-based eye-tracking: The replication of the cascade effect, the novelty preference, and the visual world paradigm. *Behavior Research Methods*, 1–14.

Vos, M., Minor, S., & Ramchand, G. C. (2022). Comparing infrared and webcam eye tracking in the visual world paradigm. *Glossa Psycholinguistics*, *1*.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations.* Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors).* Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2021b). *Tidyr: Tidy messy data.* Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation.* Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data.* Retrieved from https://CRAN.R-project.org/package=readr

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

Yarbus, A. L. (1967). *Eye movements and vision.* Plenum Press.

Zheng, C., & Usagawa, T. (2018). A rapid webcam-based eye tracking method for human computer interaction. *2018 International Conference on Control, Automation and Information Sciences (ICCAIS)*, 133–136. IEEE.