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

14      The authors made the following contributions. Joshua R. de Leeuw:

15   Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project

16   administration, Software, Supervision, Validation, Visualization, Writing - original draft,

17   Writing - review & editing; Rachel Ryskin: Conceptualization, Formal analysis,

18   Visualization, Writing - original draft, Writing - review & editing; Ariel N. James:

19   Conceptualization, Formal analysis, Visualization, Writing - original draft, Writing - review

20   & editing; Joshua K. Hartshorne: Conceptualization, Formal analysis, Visualization,

21   Writing - original draft, Writing - review & editing; Haylee Backs: Investigation,

22   Methodology, Software; Nandeeta Bala: Investigation, Methodology, Software; Laila

23   Barcenas-Meade: Investigation, Methodology, Software; Samata Bhattarai: Investigation,

24   Methodology, Software; Tessa Charles: Investigation, Methodology, Software; Gerasimos

25   Copoulos: Investigation, Methodology, Software; Claire Coss: Investigation, Methodology,

26   Software; Alexander Eisert: Investigation, Methodology, Software; Elena Furuhashi:

27   Investigation, Methodology, Software; Keara Ginell: Investigation, Methodology, Software;

28   Anna Guttman-McCabe: Investigation, Methodology, Software; Emma (Chaz) Harrison:

29   Investigation, Methodology, Software; Laura Hoban: Investigation, Methodology, Software;

30   William A. Hwang: Investigation, Methodology, Software; Claire Iannetta: Investigation,

31   Methodology, Software; Kristen M. Koenig: Investigation, Methodology, Software;

32   Chauncey Lo: Investigation, Methodology, Software; Victoria Palone: Investigation,

33   Methodology, Software; Gina Pepitone: Investigation, Methodology, Software; Margaret

34   Ritzau: Investigation, Methodology, Software; Yi Hua Sung: Investigation, Methodology,

35   Software; Lauren Thompson: Investigation, Methodology, Software.

36      Correspondence concerning this article should be addressed to Joshua R. de Leeuw,

37   124 Raymond Ave, Poughkeepsie, NY 12604, USA. E-mail: jdeleeuw@vassar.edu

Abstract

We tested whether the results of five different eye-tracking experiments in the cognitive psychology literature would replicate in a webcam-based eye-tracking 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 & 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. 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

83  drawn about human cognition.

84      Advances in software for online data collection (de Leeuw, 2015; Papoutsaki et al.,

85  2016) have the potential to address this shortcoming by expanding access to eye-tracking

86  technology for researchers and making it feasible to broaden and diversify the participant

87  samples. In particular, `Webgazer.js` (Papoutsaki et al., 2016) is a webcam-based

88  Javascript plug-in that works in the browser. It can be integrated with any Javascript web

89  interface. As a result, it can be used in conjunction with many existing platforms for online

90  behavioral data collection that are familiar to cognitive scientists, such as `jsPsych` (de

91  Leeuw, 2015), `Gorilla` (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020), or

92  `lab.js` (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2021). However, the added

93  convenience comes at the cost of spatial and temporal resolution. The extent of this loss of

94  precision and its impact on the kinds of research questions that webcam eye-tracking is

95  appropriate for are the subject of the current study.

96      A few previous studies have used webcam eye-tracking in the context of

97  behavioral/cognitive science experiments and reported on the quality of the data. In what

98  follows, we provide a brief review of the published work that we are aware of (note that we

99  focus on studies with adult participants as the considerations for eye-tracking of children

100 are substantially different, but see e.g., Bánki, Eccher, Falschlehner, Hoehl, & Markova,

101 2022).

102     Semmelmann and Weigelt (2018) compared eye-tracking results between in-lab

103 (n=29) and online (n=28) studies in three tasks: fixation, pursuit, and free viewing. They

104 used the `Webgazer.js` library (Papoutsaki et al., 2016) and programmed tasks in

105 HTML/Javascript directly. The first two tasks tested measurement of basic gaze

106 properties. In the fixation task, participants were asked to fixate a dot, and in the pursuit

107 task, they were asked to follow a dot with their eyes. In the free viewing task, the eye

108 movements may have been more semantically driven: participants were shown a picture of

a human face and asked to look wherever they wanted to on the image. Webgazer was successful in accurately detecting fixation locations and saccades in all three tasks, though online data (collected through a crowdsourcing platform) were slightly more variable and delayed relative to the in-lab data. The free viewing task replicated previously observed statistical patterns (e.g., more fixations on eyes than mouths).

Similarly, Yang and Krajbich (2021) used the `Webgazer.js` library (Papoutsaki et al., 2016) combined with the `jsPsych` library for conducting behavioral experiments in a web browser (de Leeuw, 2015), to replicate a well-established link between value-based decision-making and eye gaze Krajbich, Armel, & Rangel (2010). Online participants (n=38) first saw a series of images of 70 snack foods and rated how much they liked each one. During the primary task, on each of 100 trials, two of the snack food images were displayed on the left and right sides of the display and participants chose the one that they preferred while their gaze was monitored. As in previous work, participants were biased to choose the option they had spent more time looking at. The authors also implemented a code modification to address temporal delays in `WebGazer`.

Several papers have used the `WebGazer.js` library (via different interfaces) to replicate visual world paradigm studies. Slim and Hartsuiker (2022) used the `PCIbex` online experiment platform (Zehr & Schwarz, 2018) to replicate a study in which participants (n=90) listened to sentences (e.g., Mary reads a letter) while viewing four pictures displayed across the four quadrants of the screen [e.g., a letter, a backpack, a car and a wheelchair; Dijkgraaf, Hartsuiker, and Duyck (2017)]. When the verb in the sentence was *constraining* (e.g., read) with respect to the display (only the letter would be an appropriate continuation), listeners made more fixations to the target image (letter) than when the verb was *neutral* (e.g., Mary steals a letter), as in previous work. However, they observed a substantial delay of ~200ms in the onset of the effect relative to the in-lab study with an infrared eyetracker and a reduction in effect size (despite a threefold increase in sample size relative to the original). The authors noted that delay was particularly

problematic given that the purpose of the original study was to capture aspects of predictive or anticipatory processing: in the original study, but not in the web replication, the difference between *constraining* and *neutral* conditions emerged before the onset of the final noun (e.g., letter). Similarly, Degen, Kursat, and Leigh (2021) reported that effects of scalar implicature processing (comparing looks to the target in a four quadrant display for sentences such as "Click on the girl that has *some* apples" vs. "Click on the girl that has *three* apples" ) were smaller and delayed relative to those observed in the lab (Sun & Breheny, 2020).

In contrast, using the `Gorilla` experiment platform (Anwyl-Irvine et al., 2020), Prystauka, Altmann, and Rothman (2023) observed robust effects of verb semantics and lexical interference in similar time windows to previous in-lab studies (Altmann & Kamide, 1999; Kukona, Cho, Magnuson, & Tabor, 2014), though a direct comparison was not possible because their studies used different materials than the in-lab experiments they were conceptually replicating. Furthermore, using `jsPsych`, Vos, Minor, and Ramchand (2022) closely replicated the magnitude and timecourse of effects in a lab-based visual world paradigm where the regions of interest (ROIs) consisted of the left and right halves of the display, suggesting that the concerns about poor temporal resolution can be mitigated. Indeed, the eye-tracking plug-in in `jsPsych` uses a fork of `Webgazer.js` in which certain modifications have been made to minimize processing time. [1]

Finally, Van der Cruyssen et al. (2023) used `WebGazer.js` in conjunction with `PsychoJS` to conceptually replicate three paradigms, one which is analogous to Yang and Krajbich (2021), another testing the novelty preference (tendency for learners to look more at an object that they have not yet studied), and a visual world paradigm study with 4 objects, similar to Prystauka et al. (2023). These replications were successful but they noted reductions in effect sizes relative to an in-lab dataset. They did not address whether

---

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

161 there were temporal delays.

162     In sum, webcam eye-tracking has been used to varying degrees of success across a
163 small set of paradigms that are used in cognitive science research. One configuration,
164 `jsPsych` with a modification of `WebGazer.js`, appears to have circumvented initial
165 limitations in temporal precision (Krajbich et al., 2010; Vos et al., 2022) but has only been
166 tested with paradigms with minimal requirements in terms of spatial precision (both
167 experiments used the two sides of the display as the ROIs). Further, the extent to which
168 noise in the measurement is due to the software and hardware limitations as opposed to the
169 difference between the in-lab and online settings remains unclear. In the current work, we
170 evaluate a broad variety of paradigms in terms of their suitability for webcam eye-tracking
171 in order to provide guidance for cognitive scientists aiming to increase and diversify
172 participation in their eye-tracking-based research.

## Present work

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

### Selection of Studies

181     Studies with large effect sizes and which are known to replicate are ideal targets for
182 further replication; otherwise, it can be difficult to distinguish a failure of the method from
183 a failure of the original study to replicate. In practice, replications (successful or otherwise)
184 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

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 studies were approved by the Vassar College

Institutional Review Board.

In addition, an in-person replication was conducted for Experiment 1. Information

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 |

210 about the sample is given in the Experiment 1 Method sections.

211     In order to have adequate statistical power and precision, we aimed for 2.5x the

212 sample size of the original experiment, following the heuristic of Simonsohn (Simonsohn,

213 2015). In Experiment 5, the original sample size was so small that we opted to collect 5x

214 the number of participants to increase precision. Because of budget and time constraints

215 we were unable to replace the data for subjects who were excluded or whose data was

216 missing due to technical failures.

217 **Equipment**

218     We used a fork of the `webgazer.js` library for webcam eye-tracking (Papoutsaki et

219 al., 2016), implemented in `jsPsych`, a Javascript library for running behavioral

220 experiments in a web browser (de Leeuw, 2015). Our fork included changes to

221 `webgazer.js` in order to improve data quality for experiments in which the precise timing

222 of stimulus onsets is relevant. Specifically, we implemented a polling mode so that gaze

223 predictions could be requested at a regular interval, which improved the sampling rate

224 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. If the second attempt also lead to low validation accuracy, participants were informed that they could not participate in the study.

## 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 4.4.0; R Core Team, 2021) and the R-packages *afex* (Version 1.3.1; Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2021), *broom.mixed* (Version 0.2.9.5; Bolker & Robinson, 2020), *dplyr* (Version 1.1.4; Wickham, François, Henry, & Müller, 2021), *forcats* (Version 1.0.0; Wickham, 2021a), *ggplot2* (Version 3.5.1; Wickham, 2016), *jsonlite* (Version 1.8.8; Ooms, 2014), *lme4* (Version 1.1.35.3; Bates, Mächler, Bolker, & Walker, 2015), *lmerTest* (Version 3.1.3; Kuznetsova, Brockhoff, & Christensen, 2017), *Matrix* (Version 1.7.0; Bates & Maechler, 2021), *papaja* (Version 0.1.2; Aust & Barth, 2020), *readr* (Version 2.1.5; Wickham & Hester, 2020), *shiny* (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 1

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

274 sentence processing. For example, when looking at the display in Figure 2 and listening to

275 a sentence like "The boy will eat the. . . ," participants are more likely to look at the cake

276 than when they hear "The boy will move the. . . ," in which case they tend to look at the

277 train, presumably because cakes are edible and trains are not. Semantic information

278 available at the verb is used to anticipate upcoming linguistic input.

279 We first collected data from participants via Prolific, then conducted a follow-up

280 in-lab replication; both the remote and the in-lab replications used WebGazer and identical

281 materials and procedures.

## Method

283 All stimuli, experiment scripts, data, analysis scripts, and a pre-registration are

284 available on the Open Science Framework at https://osf.io/s82kz.

**Participants.**

***Remote sample.*** Sixty participants were paid $2.60 for their participation. Our

287 sample size of participants was determined by the total run time of our experiment, ~10

288 minutes, and the allotted funding from the Vassar College Cognitive Science Department.

289 From this information, we calculated a reasonable number of participants we could afford

290 to compensate on Prolific. Note that the sample size of the original study was 24. For

291 unknown reasons, 2 of the subjects' results were not recorded, so in the analysis, we worked

292 with data collected from 58 participants.

***In-lab sample.*** Forty-nine participants completed the study in a lab setting. They

294 were recruited via the college subject pool. Participants needed to be 18 years of age or

295 older and native speakers of English. The study was approved by the Institutional Review

296 Board at Boston College.

**Procedure.** The task began with a 9-point eye-tracker calibration and validation

298 (Figure 1). During the experiment, the participants were simultaneously presented with a

299 visual image and a corresponding audio recording of a spoken sentence. Participants had to

300 input a keyboard response indicating "yes" or "no" as to whether the sentence they heard

301 was feasible given the visual image. There were two practice trials to ensure that

302 participants understood the instructions before they undertook the main portion of the

303 experiment. Participants' reaction times, keyboard responses, and looks to objects in the

304 scene were recorded for each trial.

305 `## Warning: A numeric 'legend.position' argument in 'theme()' was deprecated in ggplot2`

306 `## 3.5.0.`

307 `## i Please use the 'legend.position.inside' argument of 'theme()' instead.`

308 `## This warning is displayed once every 8 hours.`

309 `## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was`

310 `## generated.`

311 **Materials & Design.**    The visual stimuli were created through Canva and depicted

312 an agent accompanied by four to five objects in the scene (see Figure 2). On critical trials,

313 participants heard one of two sentences associated with the scene. In the restrictive

314 condition, the sentence (e.g., "The boy will eat the cake") contained a verb (e.g., "eat")

315 which restricts the set of possible subsequent referents (e.g., to edible things). Only the

316 target object (e.g., the cake) was semantically consistent with the verb's meaning. In the

317 non-restrictive condition, the sentence (e.g., "The boy will move the cake") contained a

318 verb (e.g., "move") which does not restrict the set of possible subsequent referents. The

319 target object (e.g., the cake) as well as the distractor objects (e.g., the train, the ball, etc.)

320 were semantically consistent with the verb's meaning. Both sentences were compatible

321 with the scene, such that the correct keyboard response for the critical trials was "yes."

322 Filler trials consisted of scenes that also contained an agent surrounded by objects as in the

323 critical trials, but corresponding sentences named an object that was not present in the

324 scene. The correct keyboard response for the filler trials was "no."
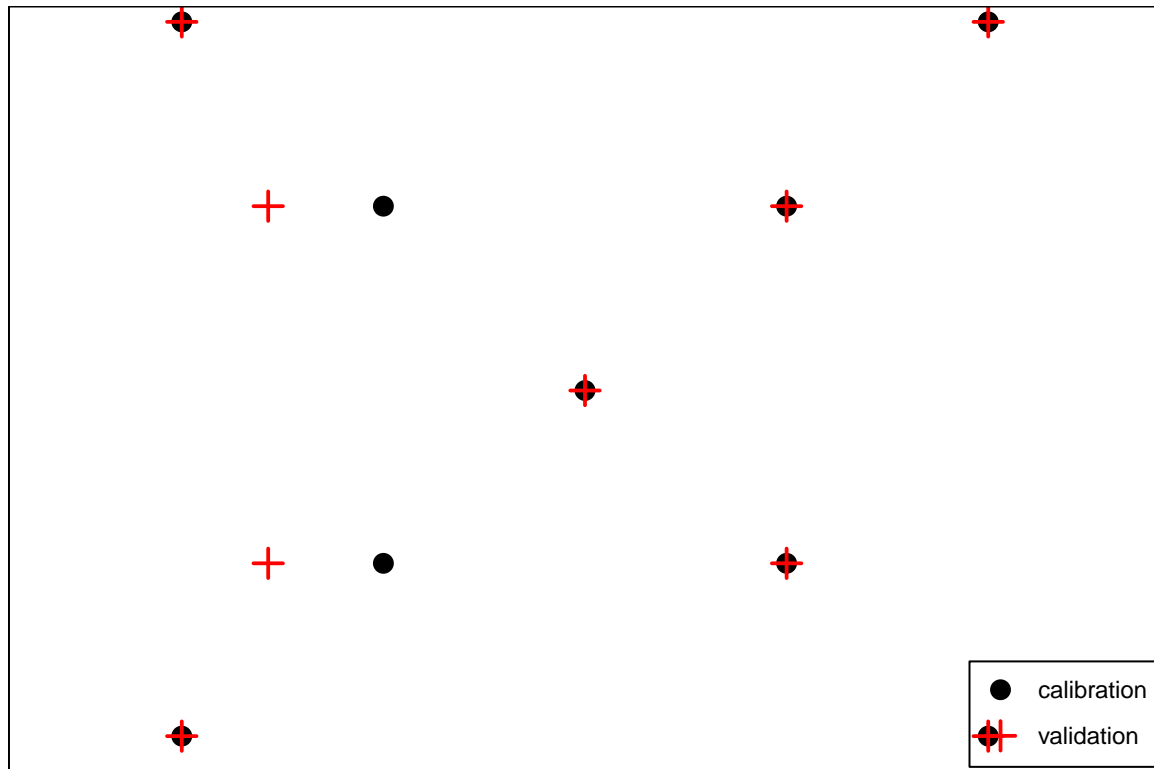
*Figure 1*. Calibration and validation point locations for Experiment 1. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.

Each participant was presented with sixteen critical trials (eight in the restrictive condition, eight in the non-restrictive condition) and sixteen 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.

**Data pre-processing and analysis.** 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.

*Figure 2*. 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.

## Results

### Remote sample.

*Minimal Exclusion.* The first set of analyses used minimal exclusion criteria. First, we eliminated participants with 0 percent of fixations in any ROIs. This resulted in the elimination of one participant. Second, we excluded participants with validation accuracy under 10 percent, resulting in an additional 5 excluded participants. The following analyses included 52 participants.

₃₄₂ ***Cumulative Fixation Probabilities.*** For each sentence, the target time window

₃₄₃ began at the onset of the verb and ended 2000 milliseconds later. This window was then

₃₄₄ divided into 50-ms bins; for each participant and each trial, we recorded whether each

₃₄₅ object was fixated during the 50-ms bin. Collapsing over trials and participants, and

₃₄₆ averaging across distractors, we calculated the cumulative probability of fixation, shown in

₃₄₇ Figure 3, Panel (b).

₃₄₈ ***Pre-noun fixations.*** In our first two analyses, we asked whether participants

₃₄₉ looked more to the target than to the distractor during the predictive time window, given

₃₅₀ that the verb is restricting. The first model tested whether there were more fixations to the

₃₅₁ target object than to the distractor in the time window before the onset of the target noun.

₃₅₂ We ran a regression model predicting the cumulative fixation probability in the last 50-ms

₃₅₃ bin before noun onset from the verb condition (restricting = 1 vs. non-restricting = 0),

₃₅₄ object type (target = 1 vs. distractor = 0), and their interaction, along with random effects

₃₅₅ for participants and images (with no covariance between random effects because the model

₃₅₆ cannot converge with full covariance matrix). There were no significant effects, although

₃₅₇ the critical interaction was in the expected direction ($b = 0.05$, $SE = 0.03$, $p=0.15$).

₃₅₈ ***Pre-verb-offset fixations.*** Altmann and Kamide tested a second model, aligning

₃₅₉ the predictive time window with the offset of the verb rather than the onset of the noun as

₃₆₀ above. When we did the same, we again saw that the critical interaction is not significant

₃₆₁ but numerically in the expected direction ($b = 0.05$, $SE = 0.03$, $p=0.20$).

₃₆₂ ***First target fixations after verb.*** Finally, we addressed whether participants

₃₆₃ looked to the target faster in the restrictive vs. the non-restrictive condition, starting after

₃₆₄ the onset of the verb. On average, participants looked to the target 349 ms after the noun

₃₆₅ onset in the restrictive condition James, Minnihan, & Watson (2023) and 452 ms after the

₃₆₆ noun onset in the non-restrictive condition James et al. (2023). Thus, first fixations were

₃₆₇ not only delayed relative to those in the previous studies compared here, but also showed a

₃₆₈ smaller difference between conditions.

369     We ran a regression model predicting the timing of the first fixation to the target

370 object, relative to the onset of the noun, with verb condition as a predictor, mean-centered

371 verb duration as a covariate, and random intercepts and condition slopes for participants

372 and scenes. There were no significant effects; participants looked sooner at the target in the

373 restrictive condition, while accounting for verb duration and its interaction with condition,

374 but this was not a statistically significant effect ($b$ = -121.91, $SE$ = 90.57, $p$=0.20).

375     ***Aggressive Exclusion.***   The second set of analyses used more aggressive exclusion

376 criteria. First, we eliminated participants with 20 percent of their fixations or fewer

377 landing in any ROIs. This resulted in the elimination of 15 participants. Second, we

378 excluded participants with validation accuracy under 50 percent, which eliminated an

379 additional 35 participants. The following analyses included 22 participants.

380     We tested the same three models under these more aggressive exclusion criteria. The

381 first two models, comparing target and distractor fixations in the predictive window,

382 produced very similar results; the critical interaction was not statistically significant

383 (Pre-noun-onset window: $b$ = 0.07, $SE$ = 0.06, $p$=0.23; Pre-verb-offset window: $b$ = 0.05,

384 $SE$ = 0.05, $p$=0.28). However, the final model, which tested the effect of verb condition on

385 saccades to the target, yielded a statistically significant result, unlike in the previous set of

386 analyses ($b$ = -193.35, $SE$ = 96.33, $p$=0.05).

387     ***Calibration.***   Participants' calibration quality was measured as the mean

388 percentage of fixations that landed within 200 pixels of the calibration point. Calibration

389 quality varied widely, ranging from 3.16% to 98.87%.

390     We tested whether a participant's calibration quality was correlated with their effect

391 size. There were three effects of interest: the verb-by-object interaction in predicting

392 fixation probabilities, both in the (1) pre-noun-onset and (2) pre-verb-offset windows

393 (calculated as the difference in target-over-distractor preference between verb conditions),

394 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 $r = 0.04$, $p = 0.78$. In the
remote sample, we saw that calibration quality was correlated with target advantage in the
pre-noun window in the restricting condition alone. This was not the case in the in-lab
sample (Pearson's $r = 0.21$, $p = 0.14$).

**In-lab sample.**

***Minimal Exclusion.*** As in the remote sample, we checked whether there were
participants with 0 percent of fixations in any ROIs and there were none. We then
excluded participants with validation accuracy under 10 percent, resulting in 2 excluded
participants. The following analyses included 47 participants.

***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 3, Panel (c).

***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 397 ms after the noun onset in the restrictive condition and 376 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 ($b = 21.70$, $SE = 115.32$, $p=0.85$). Effects of verb duration and its interaction with condition were also not statistically significant (duration: $b = $ -0.58, $SE = 0.55$, $p=0.30$; interaction: $b = $ -0.49, $SE = 0.96$, $p=0.61$ )

**Calibration.** As before, participants' calibration quality was measured as the mean percentage of fixations that landed within 200 pixels of the calibration point. Calibration

447 quality ranged from 5.13% to 97.89%.

448      We tested whether a participant's calibration quality was correlated with their effect

449 size. Across the three condition effects of interest, calibration quality was not significantly

450 correlated (pre-noun: Pearson's $r$ = -0.24, $p$ = 0.11; pre-verb-offset: Pearson's $r$ = -0.19, $p$

451 = 0.20; first fixation: Pearson's $r$ = -0.12, $p$ = 0.41. However, when the two interaction

452 effects are calculated as the target advantage in the restricting condition only (i.e. rather

453 than a difference of differences), we see a significant correlation between target advantage

454 and calibration quality in the wider pre-noun window (Pearson's $r$ = -0.16, $p$ = 0.29).

455      ***Aggressive Exclusion.***   The second set of analyses used more aggressive exclusion

456 criteria. First, we eliminated participants with 20 percent of fixations or fewer landing in

457 any ROIs. This resulted in the elimination of 7 participants. Second, we excluded

458 participants with validation accuracy under 50 percent, which eliminated an additional 23

459 participants. The following analyses included 26 participants.

460      Across all three models, results were in line with analyses using the minimal exclusion

461 criteria; none of the critical effects were statistically significant, nor were they in the

462 expected direction (Pre-noun-onset window, verb x object interaction: $b$ = -0.08, $SE$ =

463 0.06, $p$=0.15; Pre-verb-offset window, verb x object interaction: $b$ = -0.07, $SE$ = 0.05,

464 $p$=0.16; Verb effect on first target fixation: $b$ = 6.40, $SE$ = 97.92, $p$=0.95).

465 **Discussion**

466      Overall, the results of Experiment 1 paint a sobering picture of web-based

467 eye-tracking. In the remote sample, results were in the expected direction but effects were

468 smaller and delayed relative to previous work and failed to reach statistical significance; the

469 one exception was in comparing first target fixation latencies across verb conditions, but

470 only when aggressive exclusion criteria were applied to participants (INSERT WHAT

471 THOSE CRITERIA WERE). To test whether this could be explained by the variability in
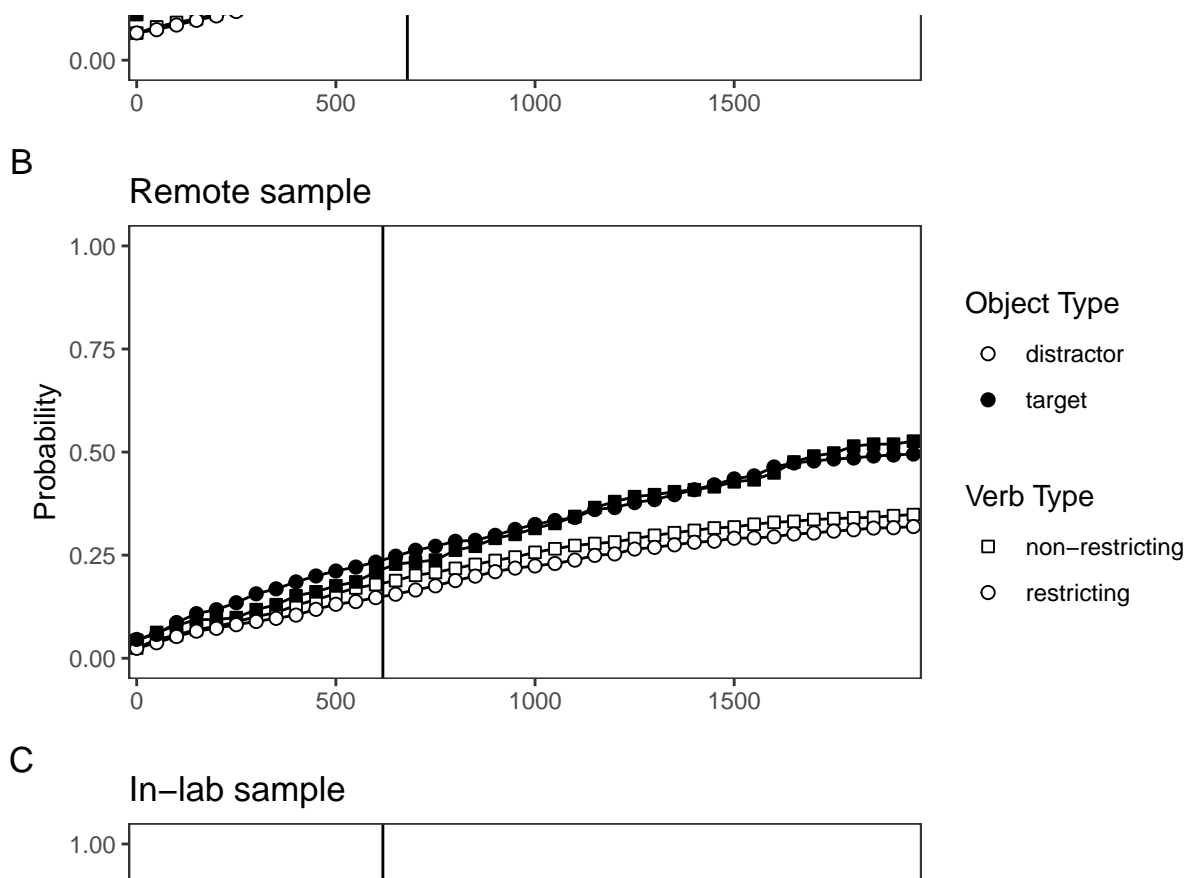
*Figure 3*. Cumulative probability of fixating distractor and target objects across conditions over time, with 0 ms aligned to the verb onset time. The vertical line marks the mean noun onset time across trials and conditions.

472 experimental settings across participants, we replicated our procedure in a lab setting.

473 Surprisingly, the results in the in-lab study were less aligned with previous work; critical

474 effects were not significant nor were they in the expected direction numerically.

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

476 well-suited for web-based eye-tracking. Notably, the ROIs were tightly drawn around the

477 five to six objects in each scene (larger ROIs would have led to overlapping objects) and

478 thus, analyses were unforgiving of inaccurate calibration. The next four experiments test

479 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.**    60 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 1 participant were not properly recorded due to unknown technical issues, so data from 59 participants were included in all analyses to follow.

**Procedure.**    The task began with a 9-point eye-tracker calibration and validation (Figure 4).

The experiment consisted of two blocks each composed of an encoding phase and a recall phase. During the encoding phase, participants saw a grid indicating the four quadrants of the screen. Each quadrant contained six images of items belonging to the same category (see Figure 5). The four categories were humanoids, household objects, animals, and methods of transportation. 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, they would have an additional 30 seconds to encode the name and orientation of
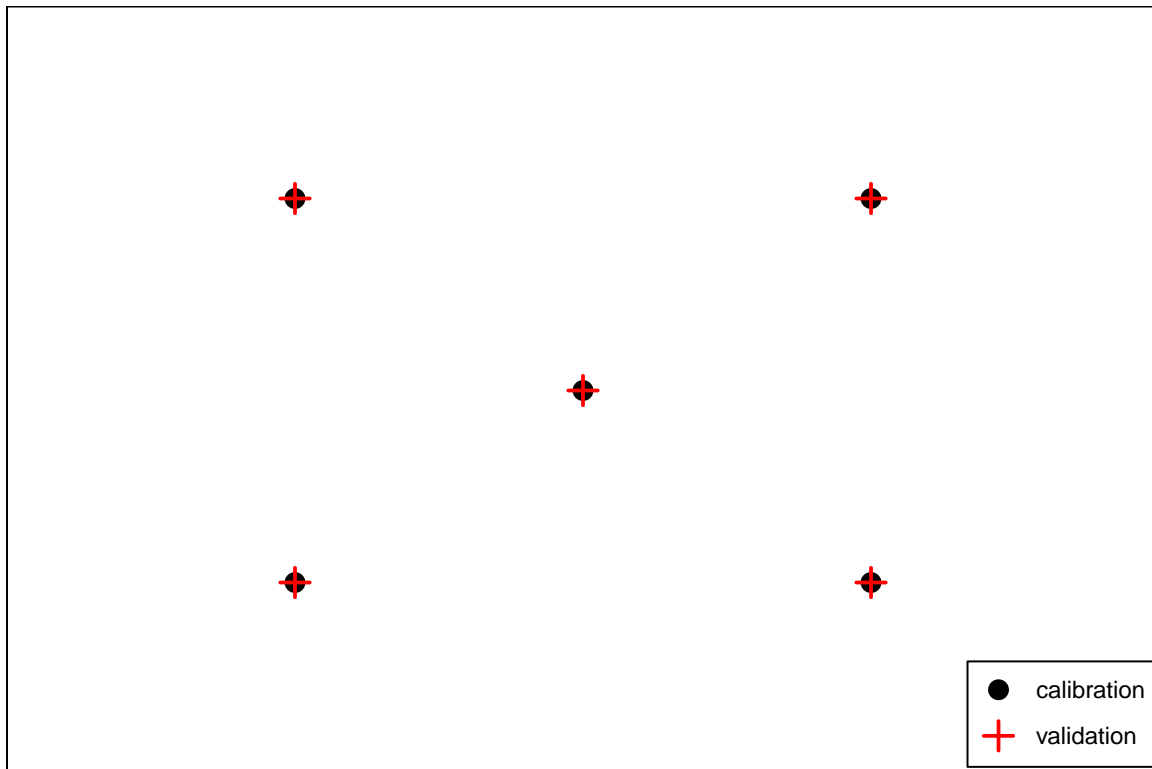
*Figure 4*. Calibration and validation point locations for Experiment 2. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.(In this experiment all the same locations were used for both calibration and validation.)

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.

During the recall phase, participants listened to statements and responded by pressing the 'F' key for false statements and 'T' for true ones. 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, with 24 interobject and 24 intraobject statements split evenly among the four quadrants. While listening to these statements, in the free-viewing block, participants saw a blank screen and were allowed to

*Figure 5*. Example trial from Experiment 2.

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 were randomly assigned to see the fixed-viewing or free-viewing block first. Different images were used in each block.

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

528 **Replication.** **Eye-gaze**. Looks during the retrieval period were categorized as

530 belonging to one of four quadrants based on the x,y coordinates. The critical quadrant was

531 the one in which the to-be-retrieved object had been previously located during encoding.

532 The other three quadrants were semi-randomly labeled "first", "second," third" (e.g., when

533 the critical quadrant was in the top left, the "first" quadrant was the top right quadrant,

534 but when the critical quadrant was in the top right, "first" corresponded to bottom right,

535 etc.). In both the fixed- and free-viewing condition, participants directed a larger

536 proportion of looks to the critical quadrant (see Figure 6). This bias appeared larger in the

537 free-viewing condition, suggesting that the manipulation was (somewhat) effective.
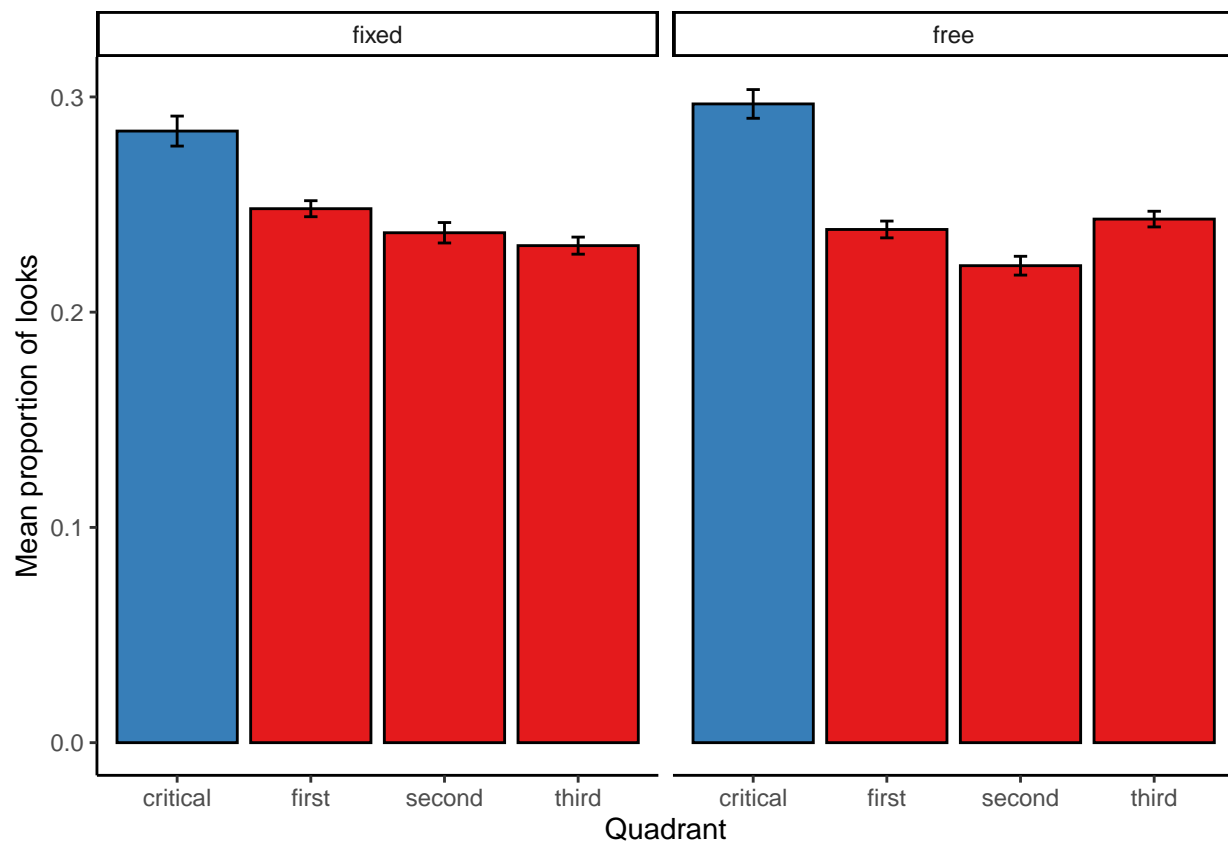


*Figure 6*. Proportion of eye-gaze to critical quadrant and other three quadrants during memory retrieval in a) fixed and b) free viewing conditions.

538      The proportions of looks across quadrants in the free-viewing condition were analyzed

539  using a linear mixed-effects model with quadrant as the predictor (critical as the reference

540  level). The model included random intercepts and slopes for participants[2]. Proportions of

541  looks were significantly higher for the critical quadrant compared to the other three (first:

542  $b$ = -0.06, $SE$ = 0.01, $p<0.001$, second: $b$ = -0.08, $SE$ = 0.01, $p<0.001$, third: $b$ = -0.05,

543  $SE$ = 0.01, $p<0.001$ )
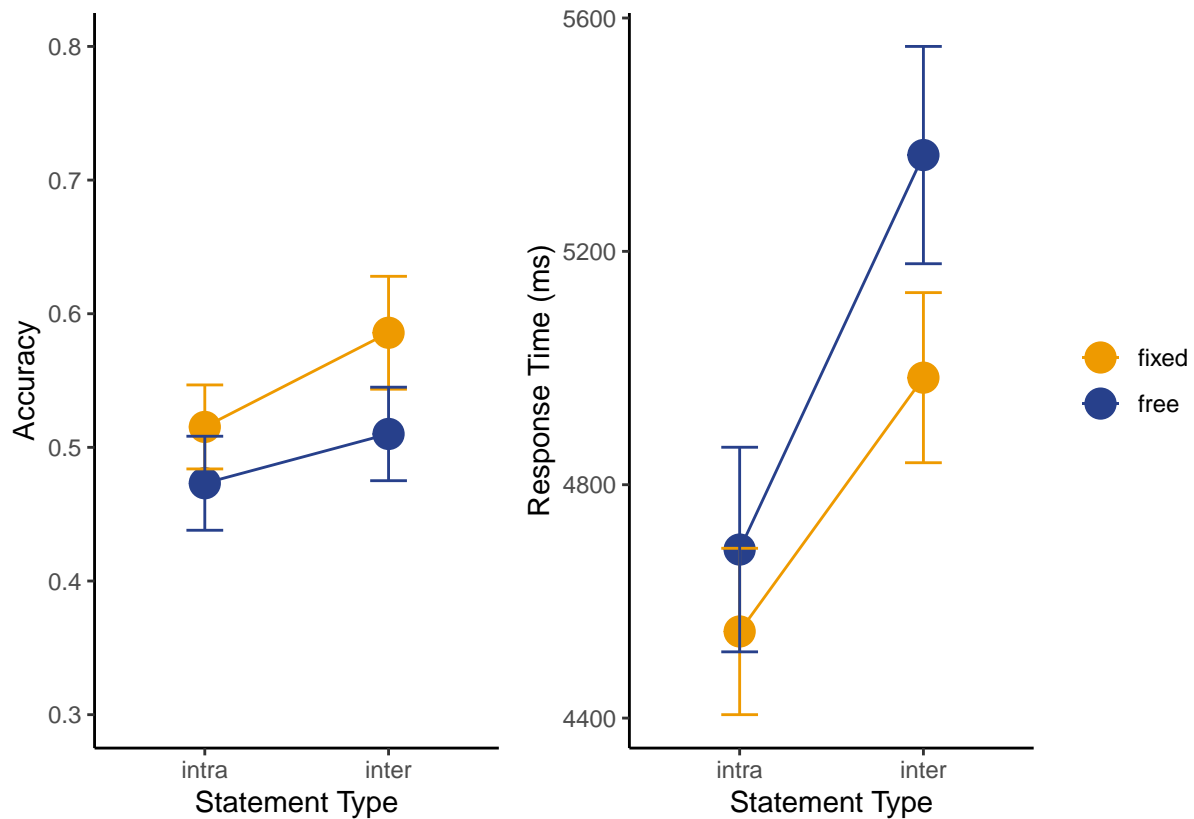


*Figure 7.* Accuracy and response times during memory retrieval.

544      **Response Time and Accuracy**. Participants' response times and accuracies on

545  memory questions are summarized in Figure 7. Both dependent variables were analyzed

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

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 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 in-lab participants 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.

**Calibration.**   Participants' calibration quality, measured as the mean percentage of fixations that landed within 200 pixels of the calibration point, varied substantially (between 17.78 and 100 %). The quality of a participant's calibration was not significantly correlated with the participant's effect size ( *Pearson's r*= 0.20, $p$ = 0.14) as measured by the difference between the proportion of looks to the critical quadrant minues the average proportion of looks to the average of the other three quadrants.

---

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

**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.

**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 8).

The experiment was administered over the course of two consecutive days. It consisted of three sections: a presentation phase, a test phase, and a recognition test. The first two phases occurred on the first day, while the recognition test occurred on the second day.

During the presentation phase, participants viewed 24 pairs of identical color photographs depicting common objects. 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.

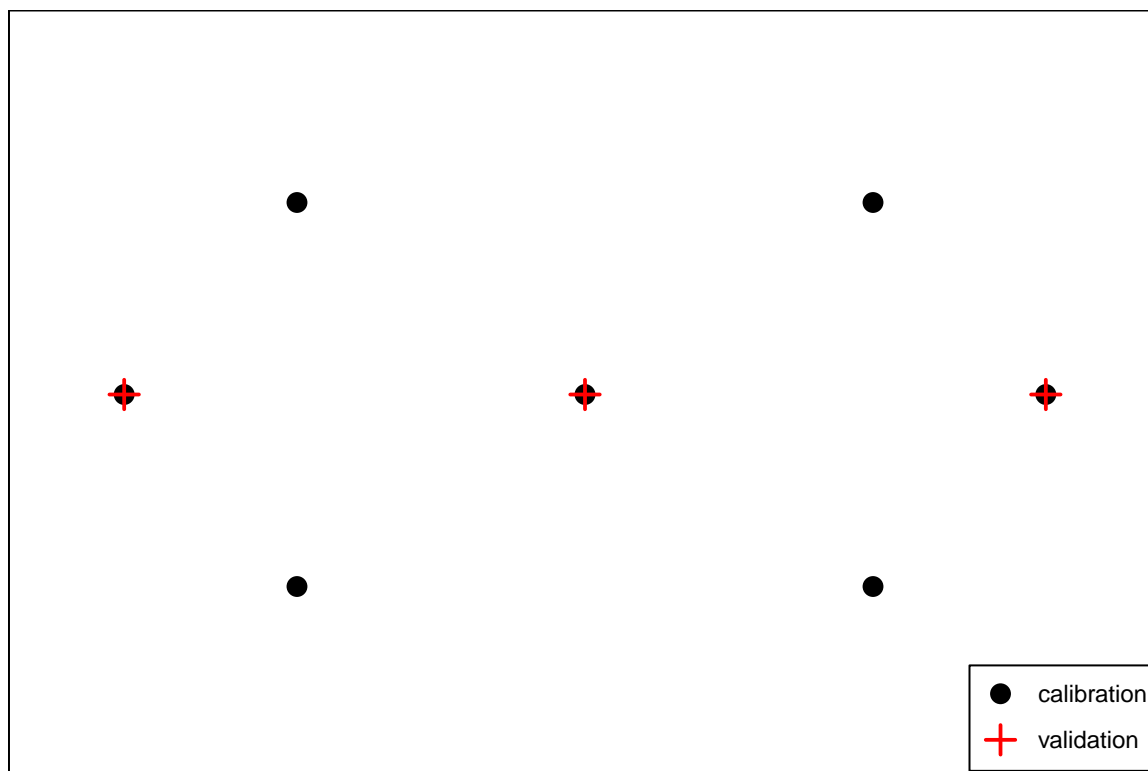After the break, they were prompted to complete the eye-tracking calibration again

*Figure 8*. 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.

before beginning the test phase. During this phase, participants again viewed 24 pairs of photographs with an interstimulus duration of 5 seconds. In each pair, one photograph was previously seen during the presentation phase, while the other was new. Which pictures were old or new was counterbalanced across participants. For half of the participants in each counterbalancing group, the new and old photographs were reversed.

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. It consisted of 48 photographs, presented one at a time. Each was shown on the screen for 1 second, followed by a 1 second interstimulus interval. Half of the photographs had been viewed twice on the previous day and were deemed the "targets." The other half depicted an object with the same name as an object in one of the old photographs, but had not been viewed before, deemed "foils." 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.

The experimental design is visually depicted in Figure 9.

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.

**Materials.**   Images were selected XXX...

**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 significantly greater than 0.5, $t(49) = 3.29$, $p = 0.00$, indicating that participants did show a preference for looking at the
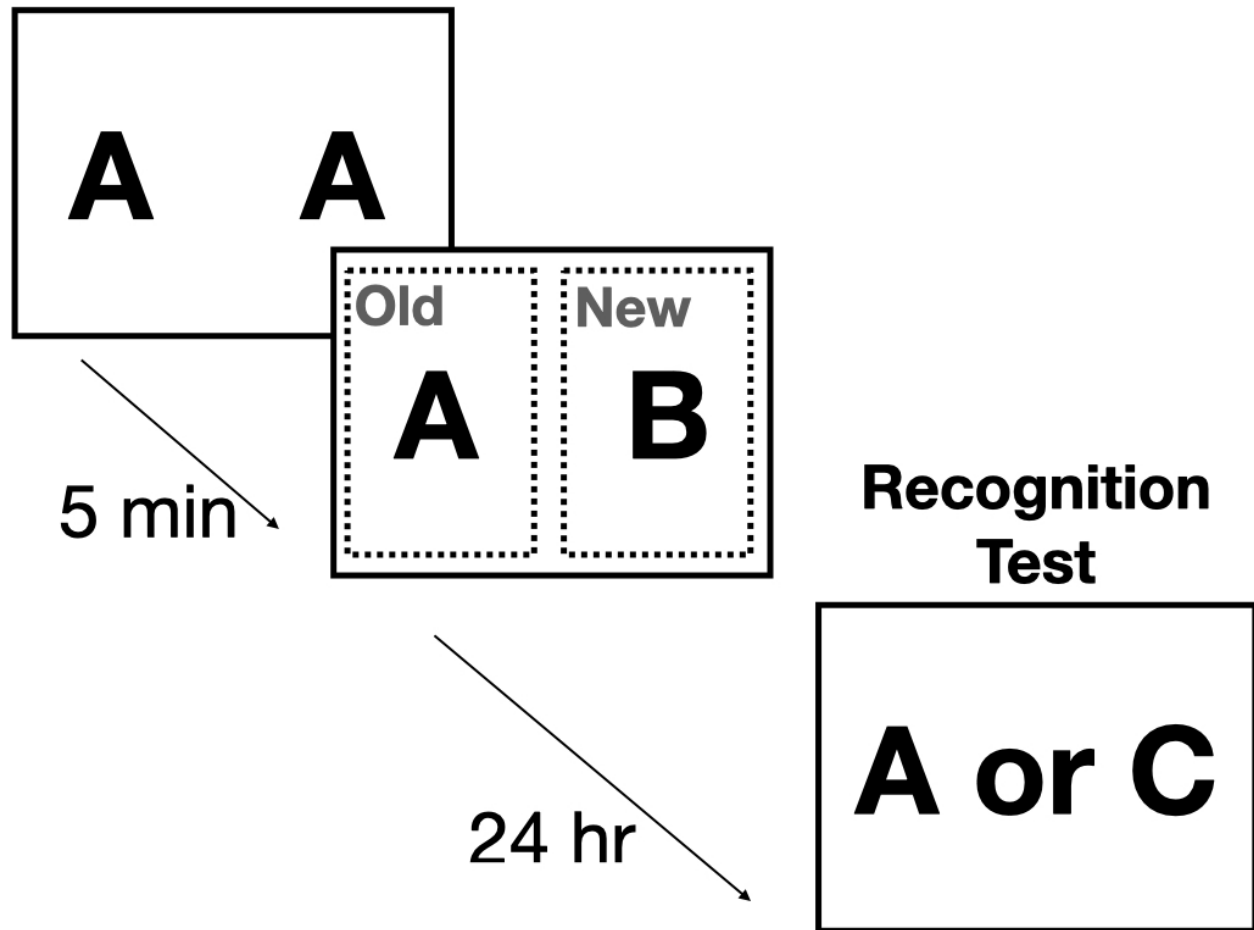
*Figure 9*. Schematic of the design of Experiment 3

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.
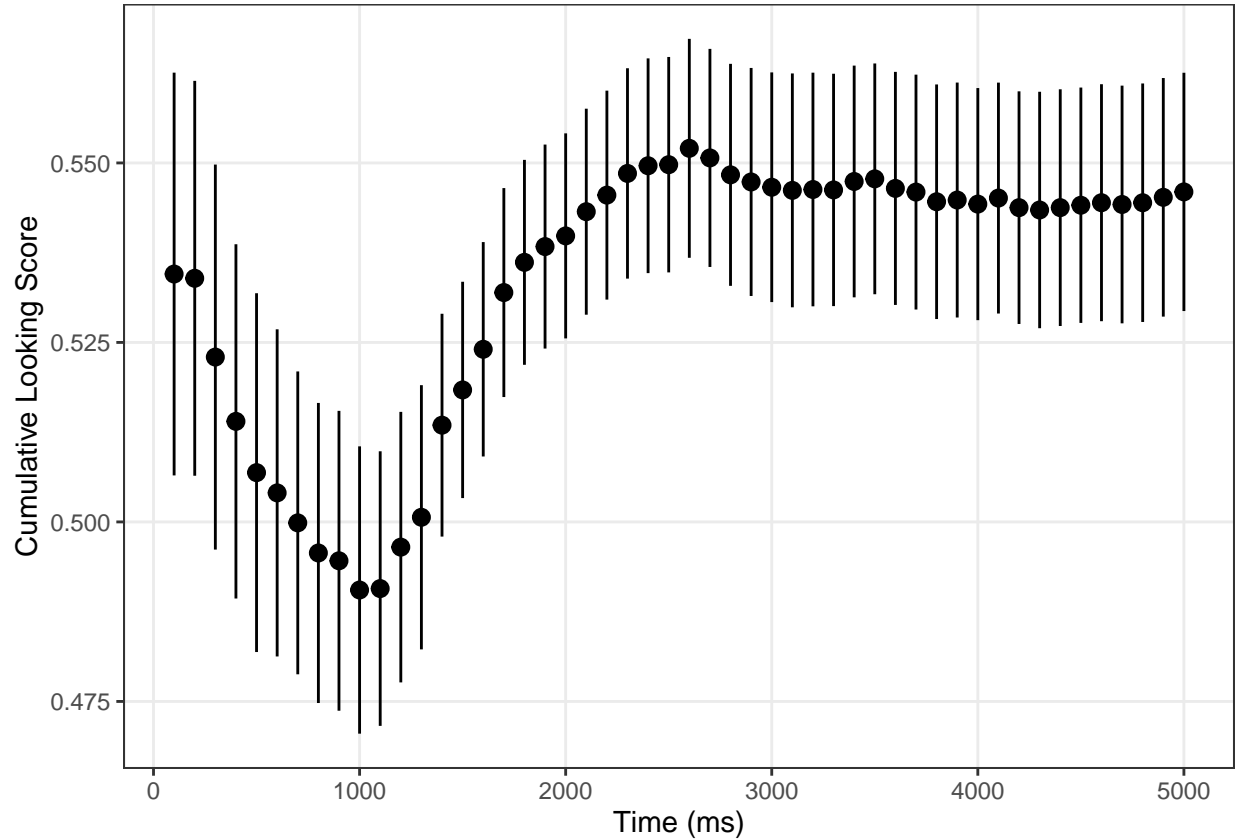
*Figure 10*. Cumulative looking score over the 5 second exposure during part 2 of day 1. Error bars represent +/- 1 SEM.

⁶⁶⁶ To see whether preferentially looking an the unfamiliar object on day 1 was

⁶⁶⁷ correlated with confidence and response time for correct responses on day 2, we computed

⁶⁶⁸ the correlation coefficient between day 1 looking scores and day 2 confidence/RT for each

⁶⁶⁹ participant. Following the original analysis, we transformed these values using the Fisher

⁶⁷⁰ p-to-z transformation. Using one-sample t-tests, we found no significant different from 0 for

⁶⁷¹ the correlation between looking score and confidence ratings, $t(38) = 0.46$, $p = 0.65$

⁶⁷² (excluding the subjects who gave the same confidence judgment for all images), nor the the

⁶⁷³ correlation between looking score and RT, $t(46) = 0.49$, $p = 0.63$.

674
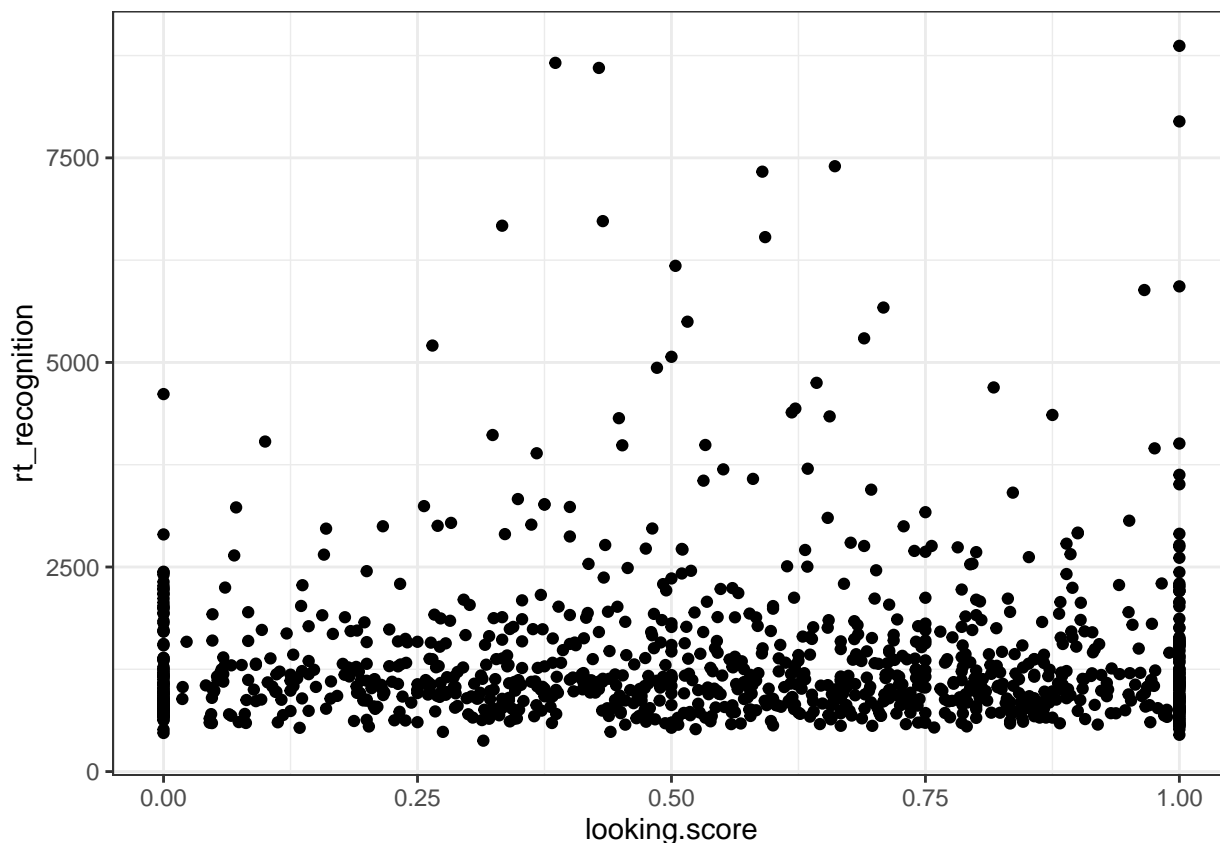
**Effects of ROIs.**    In the original experiment, the two objects on day 1 were

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.

***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, $t(50) = 3.51$, $p = 0.00$.
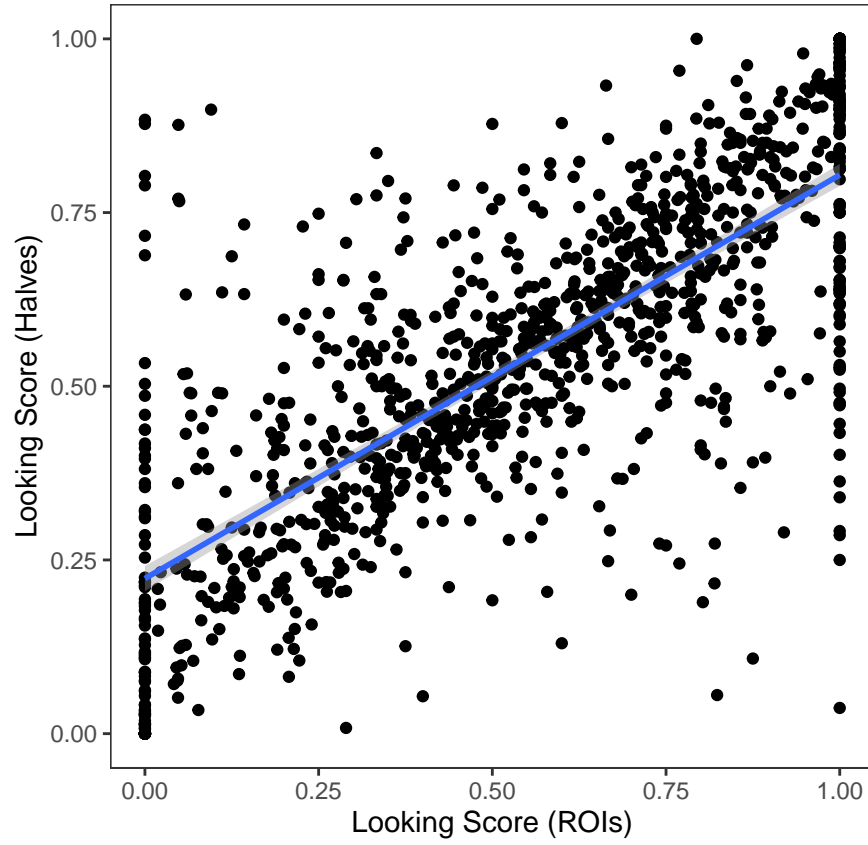
*Figure 11*. (#fig:E3-roi correlation of looking score)Correlation between looking scores calculated using ROIs and using screen halves.

686        ***Correlations with Day 2 Performance.***    Performance on day 2 remained

687    uncorrelated with day 1 looking scores after switching the coding of gaze. We found no

688    significant different from 0 for the correlation between looking score and confidence ratings,

689    t(39) = 0.74, p = 0.47 (excluding the subjects who gave the same confidence judgment for

690    all images), nor the the correlation between looking score and RT, t(47) = 0.28, p = 0.78.
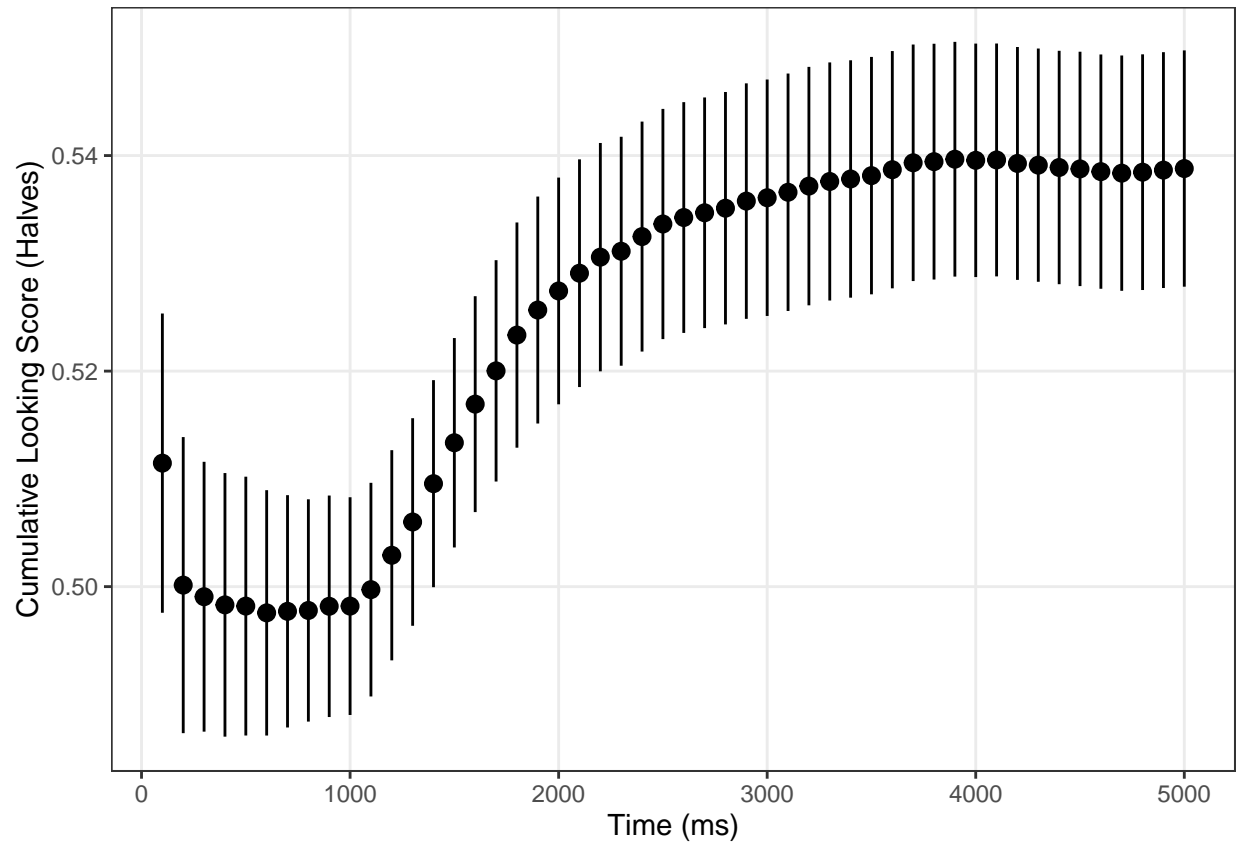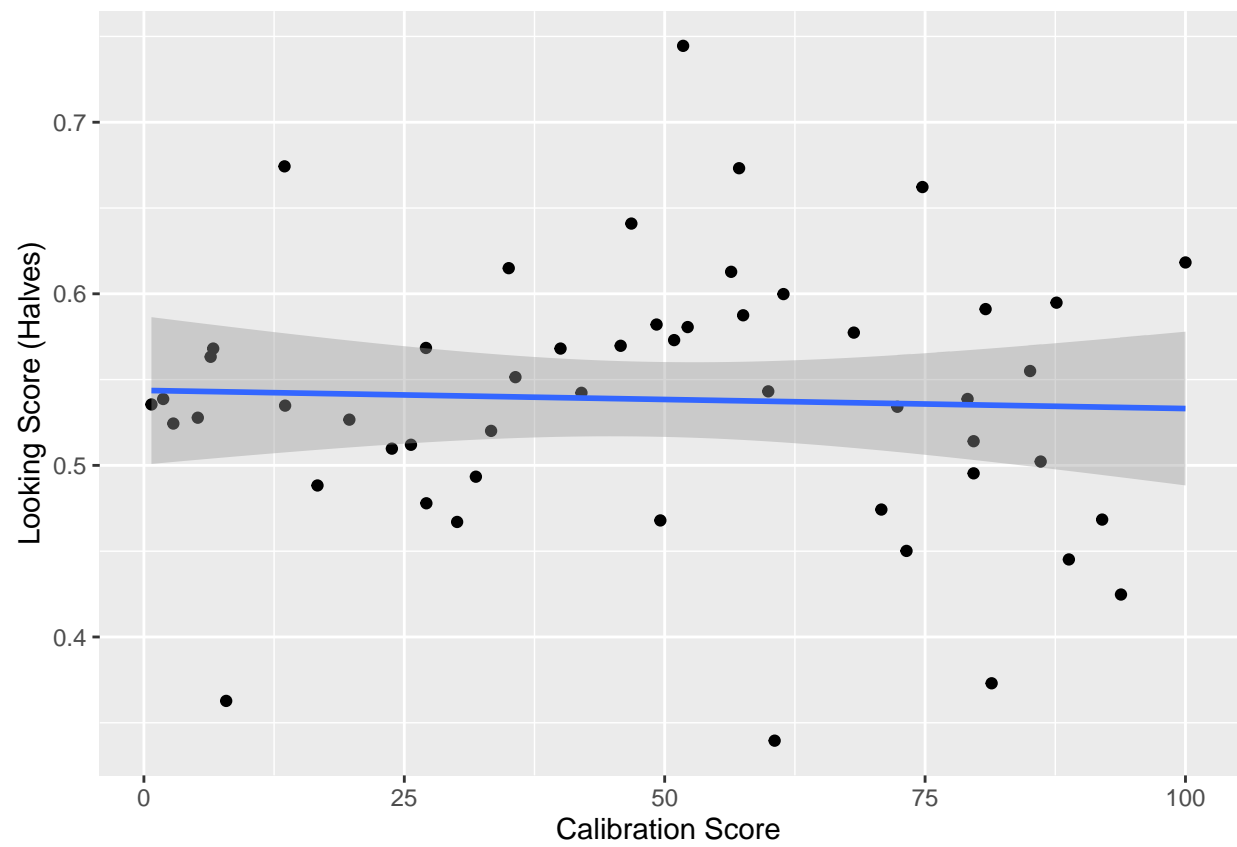
691        **Calibration.**

*Figure 12*. (#fig:E3-roi Plot of cumulative looking score)Cumulative looking score over the
5 second exposure during part 2 of day 1. Error bars represent +/- 1 SEM.

692      ***Calibration Accuracy.***



693
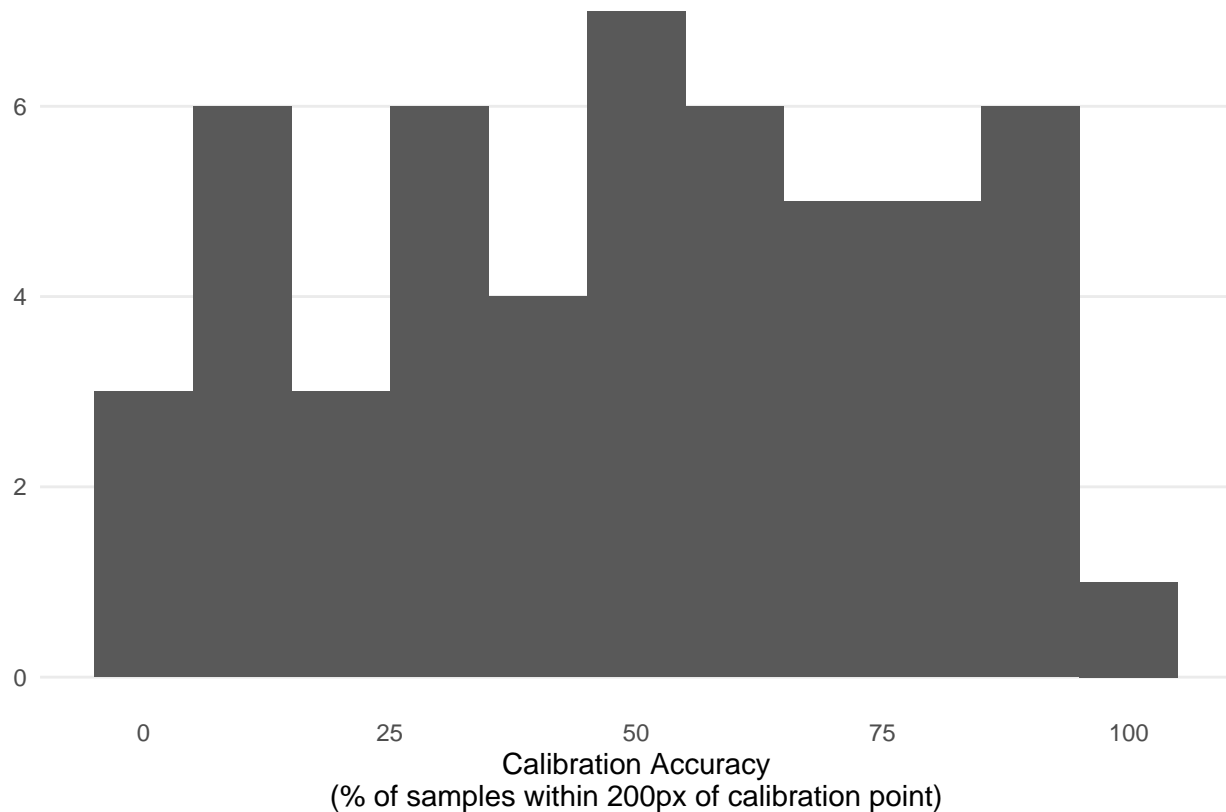
694      ***Correlation with Effects.***    To see if calibration success is correlated with the eye

695      tracking effects, we calculated a calibration score for each participant. The calibration

696      score was the average proportion of samples within 200 pixels of the validation points

697      during the final validation phase before the eye tracking is performed.

698      Calibration scores were not correlated with looking scores, regardless of which
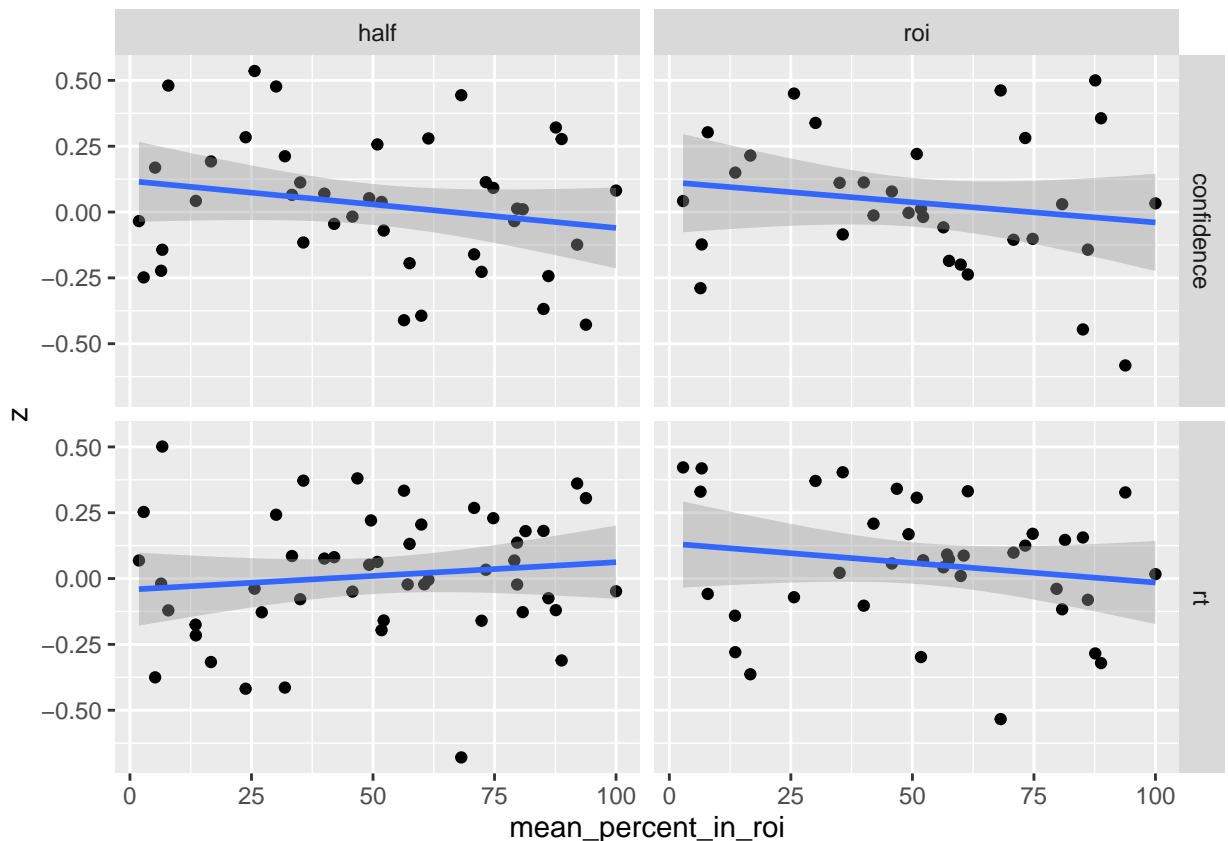
699      method was used to calculate looking scores.
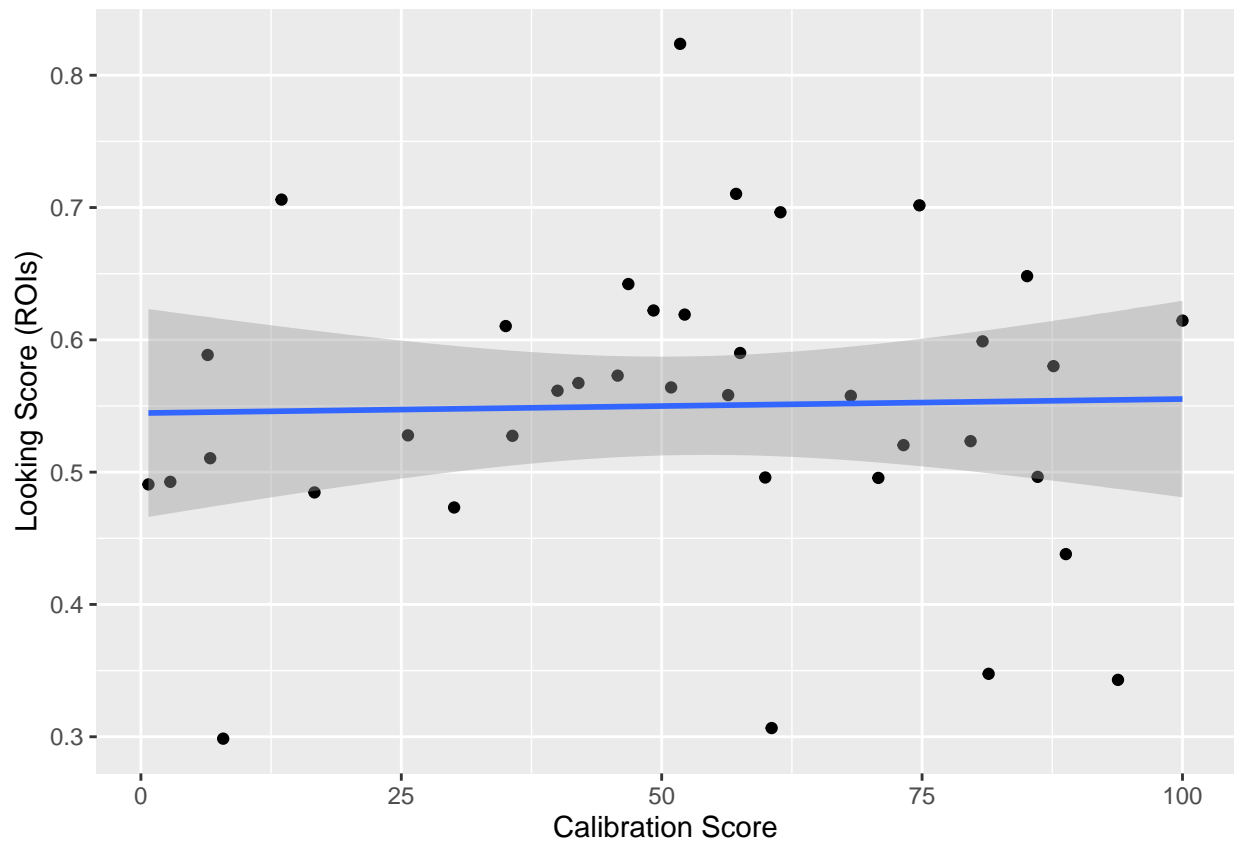
We then looked at the correlation of calibration scores with the correlation between day 2 memory performance and day 1 looking scores for both kinds of behavioral and looking measures. None of the four relationships showed a significant correlation.

705

## Discussion

706

707    As in Manns et al. (2000), participants looked more at novel images than previously

708    seen images. This effect was consistent for ROIs based on the images and for the coarser

709    ROIs based on two halves of the display. A day later, participants were also able to

710    discriminate the images they had seen from foil images they had not seen during the

711    previous session. However, there was no evidence that memory performance on day 2 was

712    related to looking time on day 1. Calibration quality did not appear to impact this

713    relationship.

## Experiment 4

714

715    The fourth study was a replication attempt of Experiment 1 in Ryskin, Qi, Duff, and

716    Brown-Schmidt (2017), which was closely modeled on Snedeker and Trueswell (2004).

717  These studies used the visual world paradigm to show that listeners use knowledge of the

718  co-occurrence statistics of verbs and syntactic structures to resolve ambiguity. For

719  example, in a sentence like "Feel the frog with the feather," the phrase "with the feather"

720  could be describing the frog, or it could be describing the instrument that should be used

721  to do the "feeling." When both options (a frog holding a feather and a feather by itself) are

722  available in the visual display, listeners rely on the verb's "bias" (statistical co-occurrence

723  either in norming or corpora) to rapidly choose an action while the sentence is unfolding. .

## Method

725  The stimuli, experimental code, and data and original analysis scripts can be found

726  on the Open Science Framework at the following link, https://osf.io/x3c49/. The

727  pre-registration for the study can be found at https://osf.io/3v4pg.

728  **Participants.**  57 participants were paid \$2.50 for their participation. A sample size

729  of 60 was initially chosen (but not reached in time) because we wanted to replicate the

730  experiment with greater statistical power. Note that the original study had a sample size of

731  24.

732  **Procedure.**  After the eye-tracking calibration and validation (Figure 13),

733  participants went through an audio test so they could adjust the audio on their computer

734  to a comfortable level. Before beginning the experiment, they were given instructions that

735  four objects would appear, an audio prompt would play, and they should do their best to

736  use their mouse to act out the instructions. They then went through three practice trials

737  which were followed by 54 critical trials and 24 filler trials presented in a random order.

738  During a trial, four pictures were displayed (target animal, target instrument,

739  distractor animal, distractor instrument), one in each corner of the screen, and participants

740  heard an audio prompt that contained instructions about the action they needed to act out
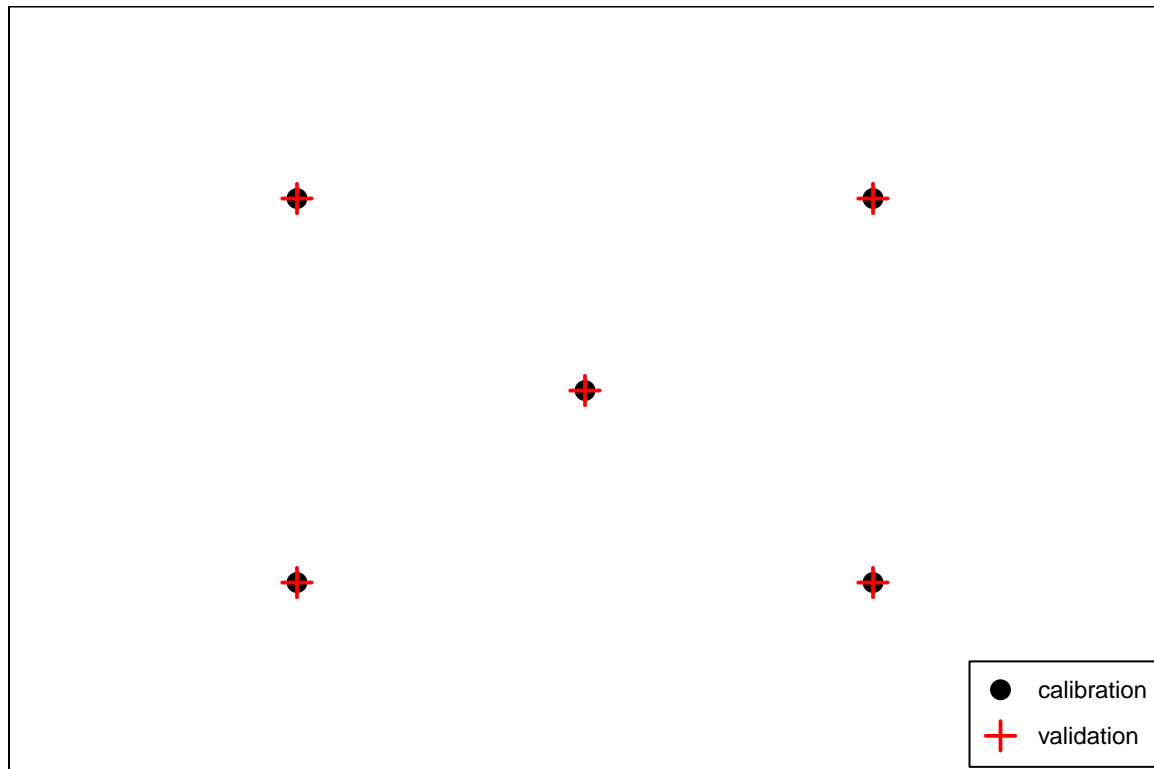
*Figure 13*. Calibration and validation point locations for Experiment 4. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.(In this experiment all the same locations were used for both calibration and validation.)

(e.g., "Rub the butterfly with the crayon"; see Figure 14)[4]. Using their cursor, participants 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

---

[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

[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.

747 and the mouse had been moved over at least one object.
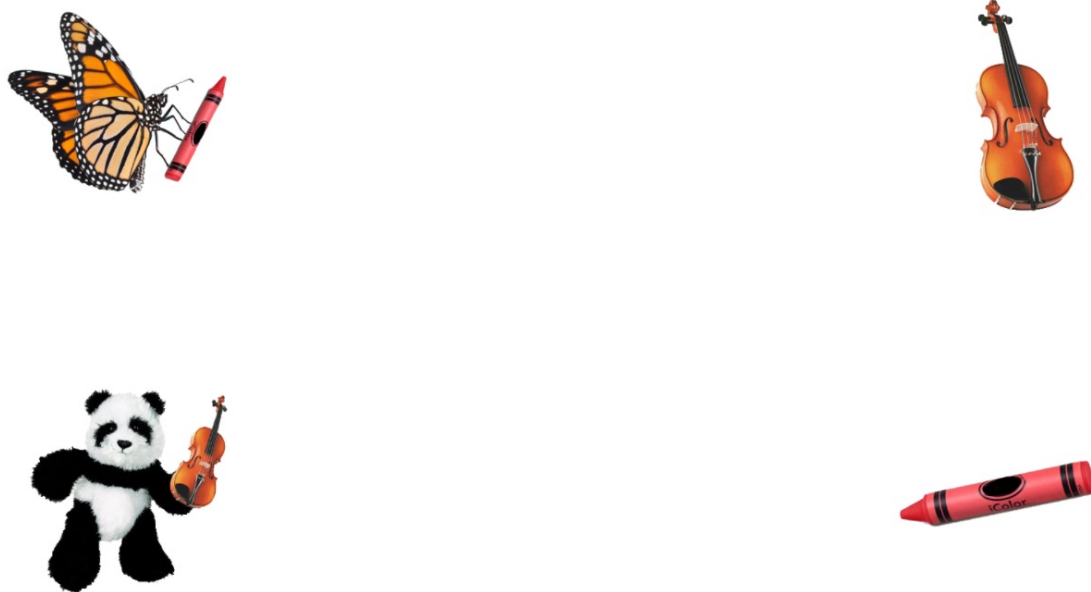


*Figure 14*. An example of a critical trial from Experiment 4 for the sentence "Rub the butterfly with the crayon." The butterfly is the target animal, the panda is the distractor animal, the crayon is the target instrument, and the violin is the distractor instrument.

748 **Materials.** The images and audios presented to the participants were the same

749 stimuli used in the original study (available here ). The critical trials were divided into

750 modifier-biased, instrument-biased, and equibiased conditions, and the filler trials did not

751 contain ambiguous instructions. Two lists of critical trials were made with different verb

752 and instrument combinations (e.g., "rub" could be paired with "panda" and "crayon" in

753 one list and "panda" and "violin" in the second list). Within each list, the same verb was

754 presented twice but each time with a different target instrument and animal. The lists were

755 randomly assigned to the participants to make sure the effects were not caused by the

756 properties of the animal or instrument images used. The list of verbs used can be found in

757 Appendix A of the original study.

<sub>758</sub> **Results**

<sub>759</sub>     **Replication.**    The location of initial mouse movements was used to assess whether

<sub>760</sub> the final interpretation of ambiguous sentences was biased by the verb. Figure 15 suggests

<sub>761</sub> that listeners were more likely to move their mouse first over the target instrument when

<sub>762</sub> the verb was equi-biased than when the verb was modifier-biased and even more so when

<sub>763</sub> the verb was instrument-biased. The opposite graded pattern can be observed for mouse

<sub>764</sub> movements over the target animal.
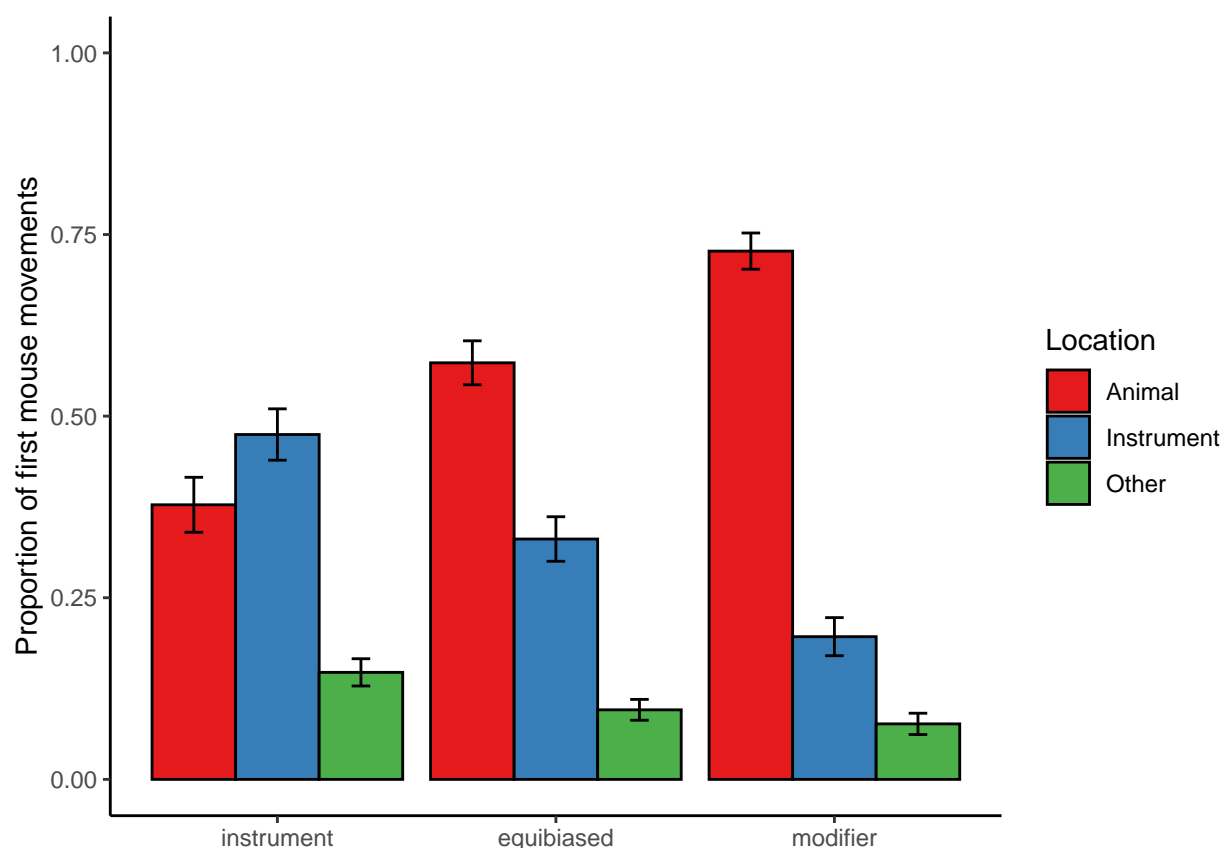


*Figure 15*. Proportion of first mouse movements by location and verb bias.

<sub>765</sub>     A mixed-effects logistic regression model was used to predict whether the first

<sub>766</sub> movement was on the target instrument with the verb bias condition as an orthogonally

<sub>767</sub> contrast-coded (instrument vs. equi & modifier: inst = -2/3, equi = 1/3, mod = 1/3; equi

<sub>768</sub> vs. modifier: inst = 0, equi = -1/2, mod = 1/2 ) fixed effect. Participants and items were

769   entered as varying intercepts with by-participant varying slopes for verb bias condition[6].

770   Participants were more likely to first move their mouse over target instruments in the

771   instrument-biased condition relative to the equi-biased and modifier-biased condition ($b =$

772   -1.50, $SE = 0.25$, $p < 0.01$). Further, participants were more likely to first move their

773   mouse over target instruments in the equi-biased condition relative to the modifier-biased

774   condition ($b =$ -1.10, $SE = 0.29$, $p < 0.01$)

775          Gaze fixations were time-locked to the auditory stimulus on a trial by trial basis and

776   categorized as being directed towards one of the four items in the display if the x, y

777   coordinates fell within a rectangle containing the image. Figure 16 suggests that the

778   participants made more fixations to the target animal when the verb was modifier-biased

779   compared to when the the verb was equi-biased and they looked at the target animal least

780   when the verb was instrument-biased. The pattern was reversed for looks to the target

781   instrument.

782          In order to assess how verb bias impacted sentence disambiguation as the sentence

783   unfolded, the proportion of fixations was computed in three time windows: the

784   verb-to-animal window (from verb onset + 200 ms to animal onset + 200 ms), the

785   animal-to-instrument window (from animal onset + 200 ms to instrument onset + 200 ms),

786   and the post-instrument window (from instrument onset + 200 ms to instrument onset +

787   1500ms + 200 ms). Mixed-effects linear regression models were used to predict the

788   proportions of fixations to the target animal within each time window with the verb bias

789   condition as an orthogonally contrast-coded (instrument vs. equi & modifier: inst = -2/3,

790   equi = 1/3, mod = 1/3; equi vs. modifier: inst = 0, equi = -1/2, mod = 1/2 ) fixed effect.

791   Participants and items were entered as varying intercepts[7]. In the *verb-to-noun* window,

_____

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

[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
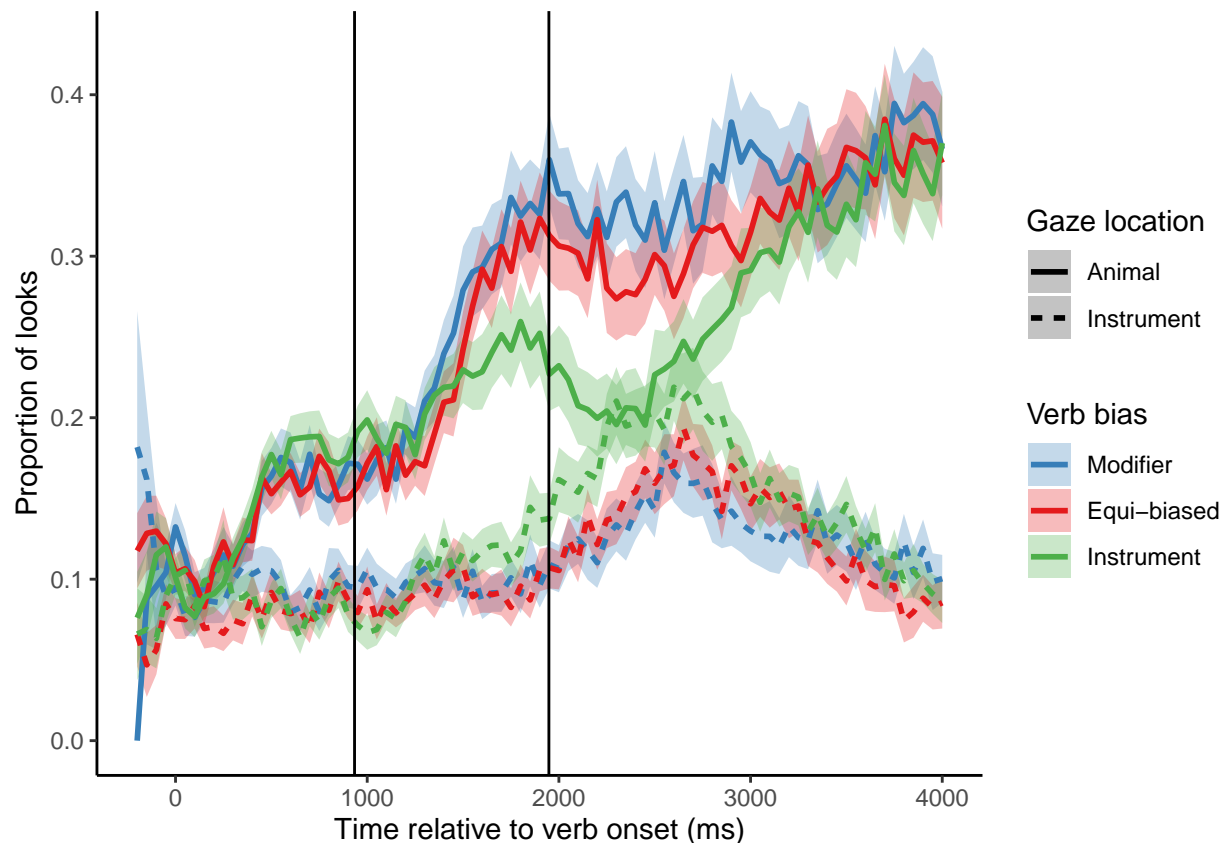
*Figure 16*. Timecourse of eye-gaze to target animal and target instrument by verb bias condition. Vertical lines indicate average onsets of animal and instrument offset by 200ms.

792 participants did not look more at the target animal in any of the verb bias conditions

793 (Instrument vs. Equi and Modifier: $b$ = -0.01, $SE = 0.02$, $p = 0.59$; Equi vs. Modifier: $b =$

794 0, $SE = 0.02$, $p = 1$ ). In the *noun-to-instrument* window, participants looked more at the

795 target animal in the modifier-biased condition and equi-biased conditions relative to the

796 instrument-biased condition ( $b = 0.03$, $SE = 0.01$, $p < 0.01$) and in the modifier biased

797 relative to the equi-biased condition ( $b = 0.02$, $SE = 0.01$, $p < 0.05$). In the

798 *post-instrument* window, participants looked more at the target animal in the

799 modifier-biased condition and the equi-biased conditions relative to the instrument-biased

800 condition ( $b = 0.08$, $SE = 0.02$, $p < 0.01$) but not significantly so in the modifier biased

—————

but did not converge.

801  condition relative to the equi-biased condition ( $b = 0.03$, $SE = 0.02$, $p = 0.15$).

802  **Comparison to in-lab data.**  The web version of the study qualitatively replicates

803  the action and eye-tracking results of the original dataset (Ryskin et al., 2017). The mouse

804  click results from both studies are summarized in Figure 17. The quantitative patterns of

805  clicks were similar to those observed in the original dataset, though for Instrument-biased

806  verbs, clicks were closer to evenly split between the animal and the instrument relative to

807  the in-lab study where they were very clearly biased toward the instrument.
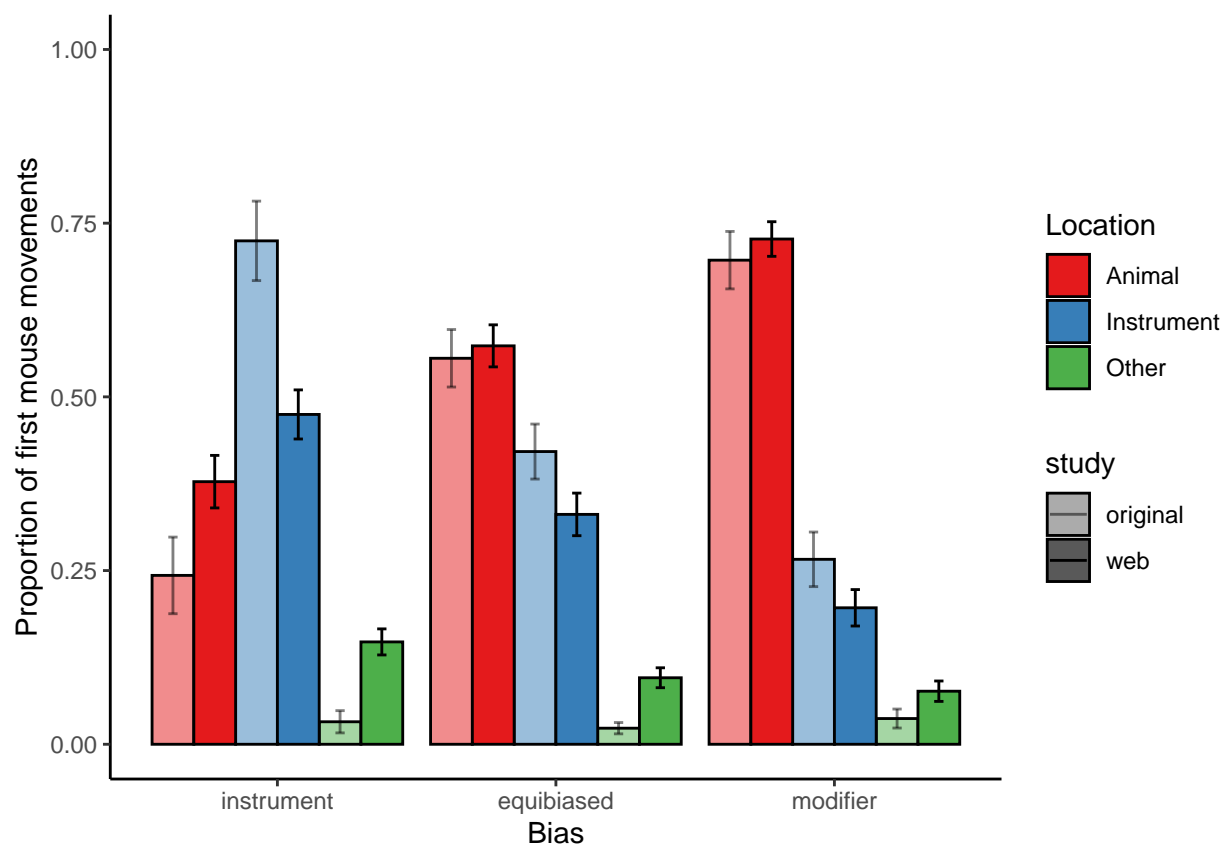


*Figure 17.* Proportion of first mouse movements by location and verb bias in the original
dataset (Ryskin et al., 2017) and the current data collected online.

808  The eye-tracking results from both studies are summarized in Figure 18. For

809  simplicity, and to reflect the dependent variable used in analyses, we average the

810  proportion of fixations to the target animal within each time window. Though the

<sup>811</sup> qualitative patterns are replicated, proportions of fixations to the target animal were much

<sup>812</sup> lower in the web version of the study. This may reflect the fact that participants in the web

<sup>813</sup> study are less attentive and/or the quality of the webgazer eye-tracking system is lower,

<sup>814</sup> relative to the Eyelink 1000 which was used for the original study.
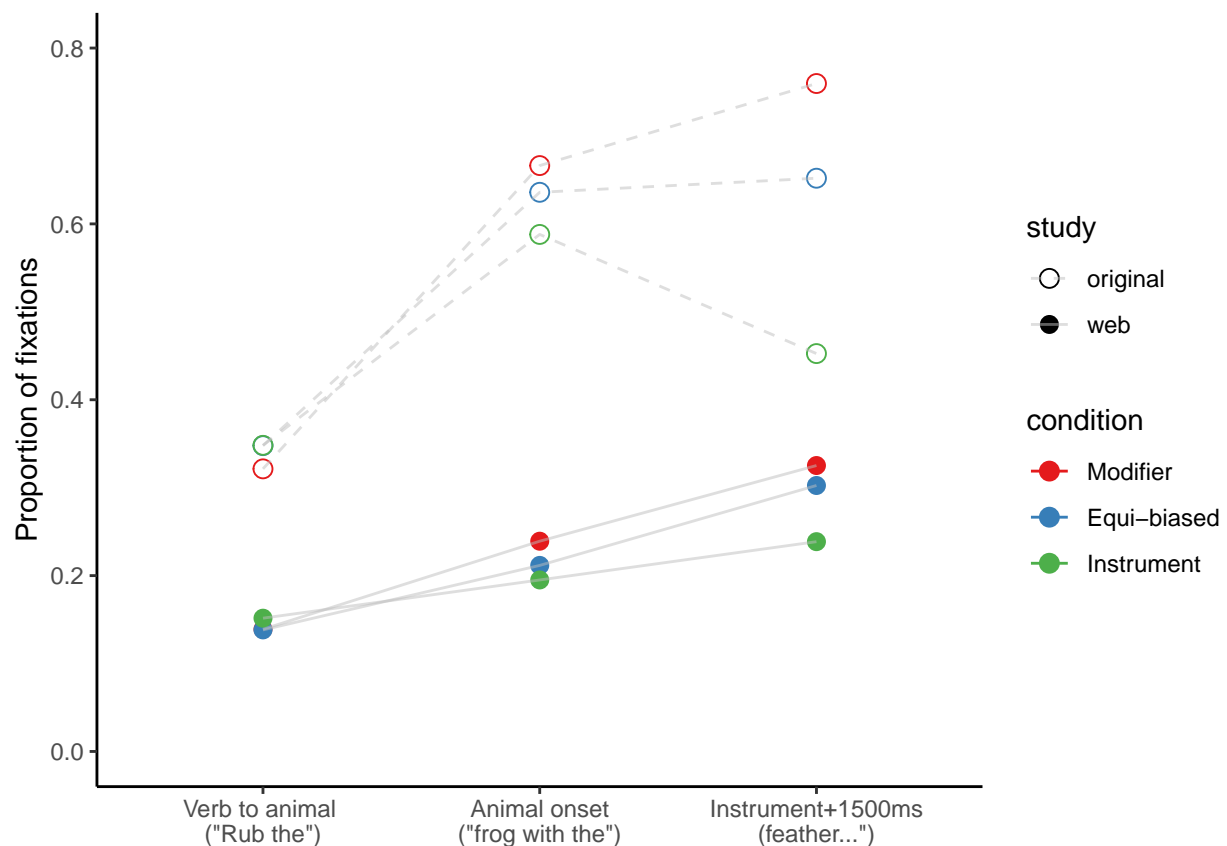


*Figure 18*. Proportion of target fixations by verb bias in the original dataset (Ryskin et al., 2017) and the current data collected online. Error bars reflect bootstrapped 95% CIs over subject means

<sup>815</sup> **Calibration.** Participants' calibration quality, measured as the mean percentage of

<sup>816</sup> fixations that landed within 200 pixels of the calibration point, varied substantially

<sup>817</sup> (between 2.22 and 97.36 %). The quality of a participant's calibration significantly

<sup>818</sup> correlated with the participant's effect size ( *Pearson's r*= 0.29, $p < 0.05$). The difference

<sup>819</sup> in target animal fixation proportions between modifier and instrument conditions was

820    higher for participants with better calibration

821          Replicating the linear mixed-effects analysis (in the post-instrument onset time

822    window only) on a subset of 35 participants with calibration quality >50% suggests that

823    the effect of verb bias condition was larger in this subset than in the full dataset.

824    Participants looked more at the target animal in the modifier-biased condition and the

825    equi-biased conditions relative to the instrument-biased condition ( $b = 0.10$, $SE = 0.02$, $p$

826    $< 0.001$) but not significantly so in the modifier biased condition relative to the equi-biased

827    condition ( $b = 0.02$, $SE = 0.02$, $p = 0.29$).

828          Replicating the linear mixed-effects analysis (in the post-instrument onset time

829    window only) on a subset of 19 participants with calibration quality >75% suggests that

830    the effect of verb bias condition was larger in this subset than in the full dataset.

831    Participants looked more at the target animal in the modifier-biased condition and the

832    equi-biased conditions relative to the instrument-biased condition ( $b = 0.11$, $SE = 0.03$, $p$

833    $< 0.001$) but not significantly so in the modifier biased condition relative to the equi-biased

834    condition ( $b = 0.05$, $SE = 0.03$, $p = 0.13$).

835          **Effects of ROIs.**   Eye-tracking on the web differs critically from in-lab eye-tracking

836    in that the size of the display differs across participants. Thus the size of the ROIs differs

837    across participants. The current version of the web experiment used a bounding box around

838    each image to determine the ROI. This approach is flexible and accomodates variability in

839    image size, but may exclude looks that are directed at the image but fall outside of the

840    image (due to participant or eye-tracker noise) as show in Figure 19a. Alternatively, The

841    display can be split into 4 quadrants which jointly cover the entire screen (see Figure 19b).

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

843    coordinates fell in, increases the overall proportions of fixations (see Figure 20). In the

844    *post-instrument* window, participants looked more at the target animal in the

845    modifier-biased condition and the equi-biased conditions relative to the instrument-biased
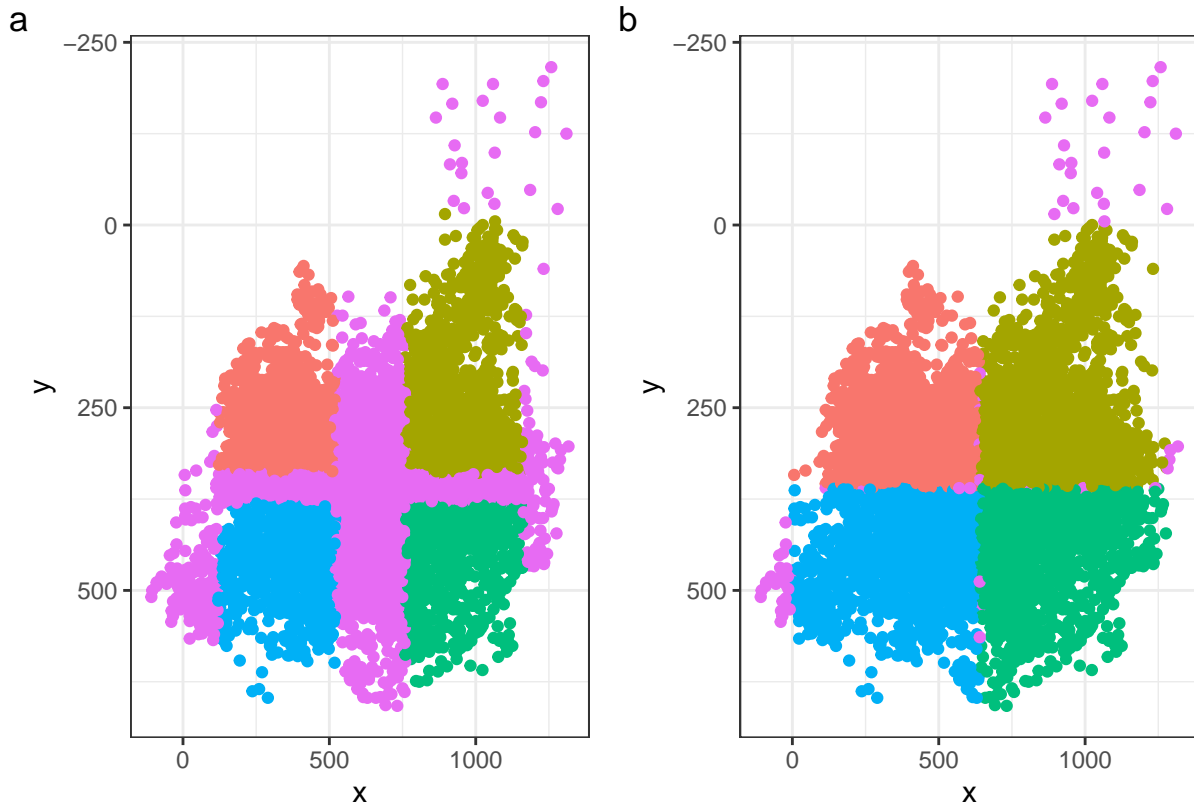
*Figure 19*. Example participant's gaze coordinates categorized into ROIs based on a) image bounding boxes and b) screen quadrants. Magenta points indicate looks that were not categorized into an ROI

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,
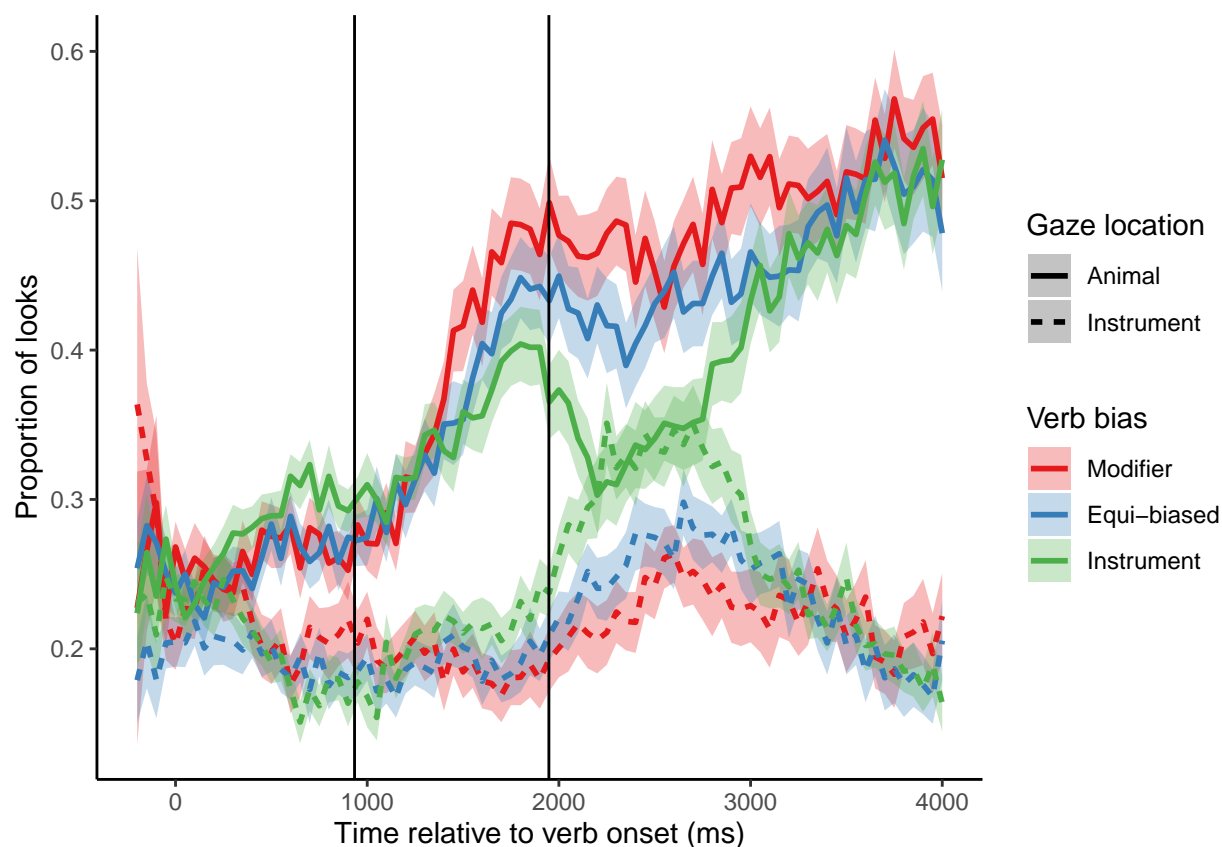
*Figure 20*. Timecourse of eye-gaze to target animal and target instrument by verb bias condition with gaze categorized based on which quadrant of the screen the coordinates fall in (as opposed to a bounding box around the image). Vertical lines indicate average onsets of animal and instrument offset by 200ms.

854 participants looked more quickly at the target animal and less at the potential instrument

855 than for instrument-biased verbs (and equi-biased verbs elicited a gaze pattern between

856 these extremes). This pattern was stronger for those who achieved higher calibration

857 accuracy and when quadrant-based ROIs were used compared to image-based ROIs.

## Experiment 5

858

859 The fifth study was a replication attempt of Shimojo et al. (2003), which found that

860 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.** 50 participants for the main task were recruited on Prolific and were paid $10/hour. 8 subjects, 4 from the attractiveness task group and 4 from the roundness task group, were excluded for incorrect validations. After this data exclusion, we ended up with 21 participants each for the attractiveness task and the roundness task. The original sample size in Shimojo et al. (2003) was 10 participants total.

**Procedure and Design.** 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 21).

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 22). Participants were randomly assigned to one of two tasks: attractiveness or shape judgment. In the attractiveness task, participants were asked to chose the more attractice face in the pair and in the shape judgment task participants were asked to pick the face that appeared rounder. They
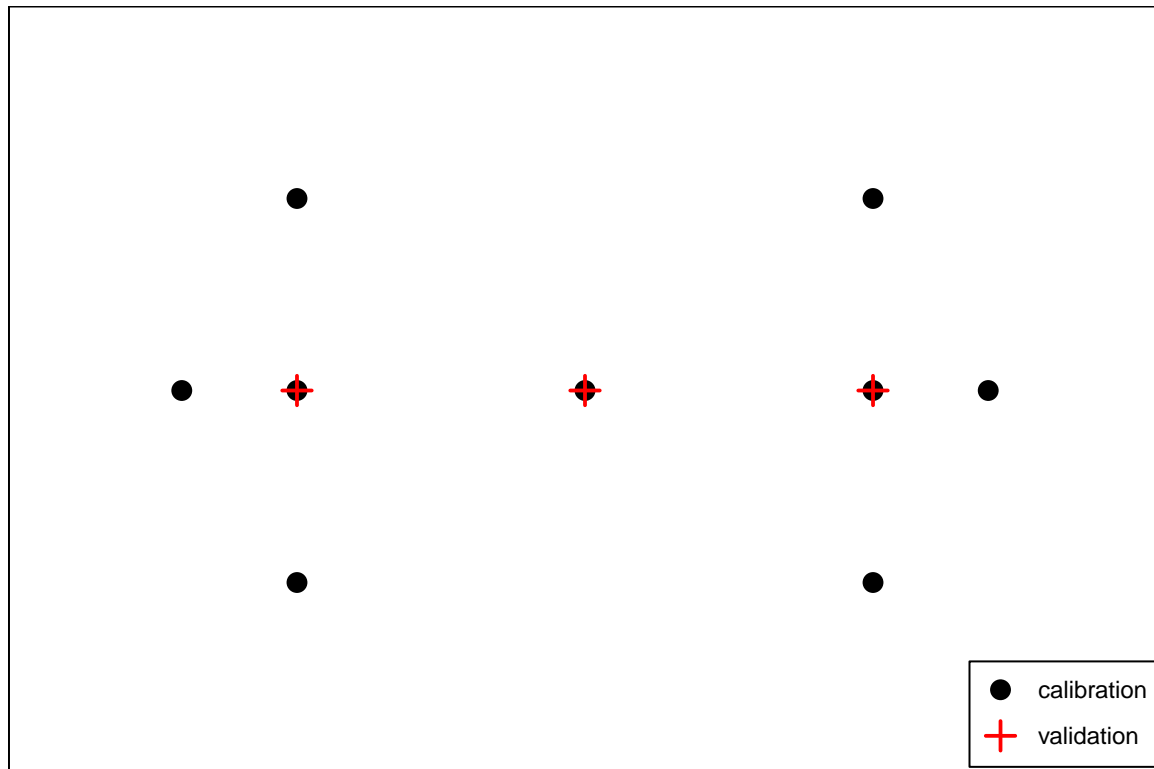
*Figure 21*. Calibration and validation point locations for Experiment 5. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.

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.

**Materials and Norming.** 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. 27 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
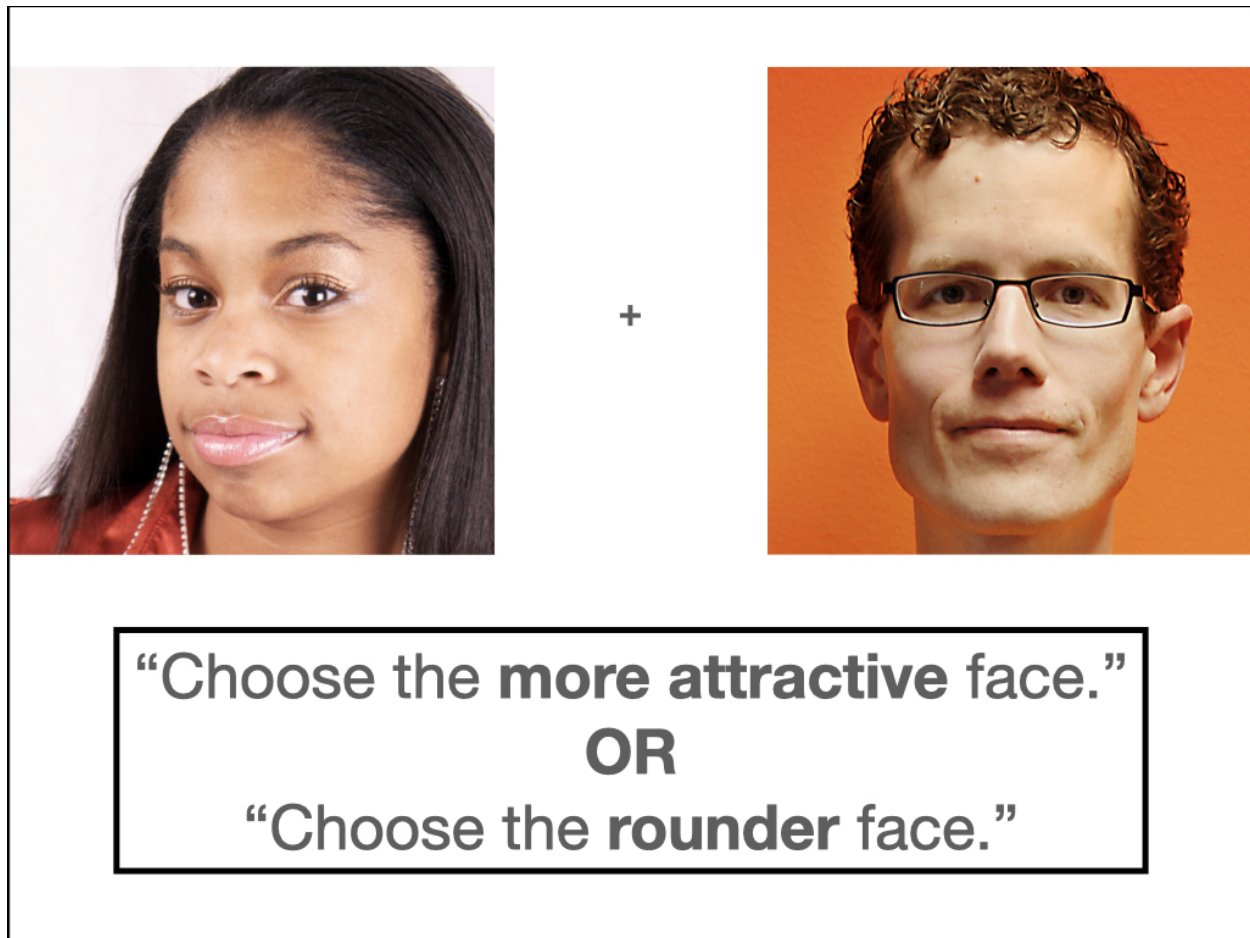
*Figure 22*. An example of a critical trial from Experiment 5. (Text did not appear on each screen.)

participant. Data from 3 participants were excluded because their mode 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).

**Data analysis.**   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 was then 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.[TODO - CONFIRM]

## Results

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 1.85 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. 23). These fit less well than in the original paper for both the attractiveness judgment ($R^2 = 0.84$ vs. 0.91) and the roundness judgment ($R^2 = 0.54$ 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 ($D = 0.19$, $p = 0.53$), 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. The distribution across participants
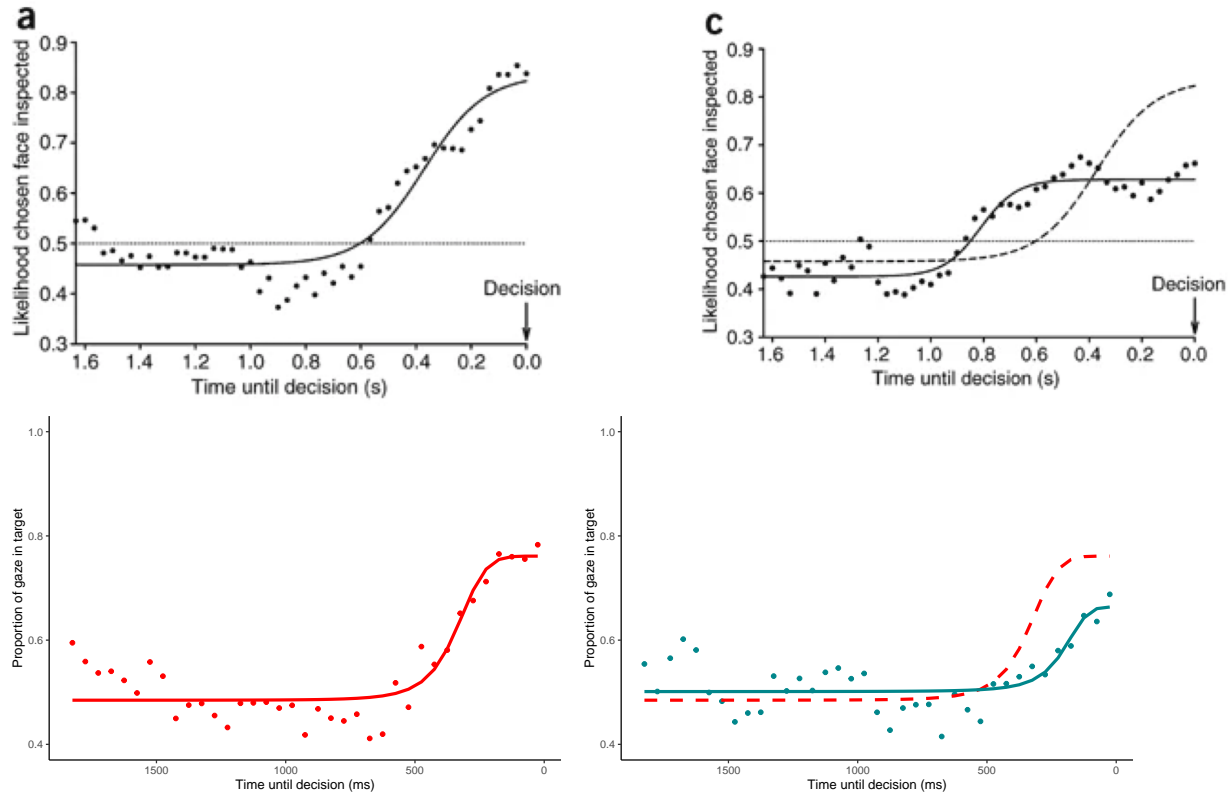
*Figure 23*. Primary results from Exp. 5. *Top* shows the original results from Shimojo and colleagues (Figures reprinted with permission[TODO]). The attractiveness judgment along with the best-fitting sigmoid is shown in the *top left*. Results for the roundness judgment are show in the *top right*, 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.

932  is shown in Fig. 24.
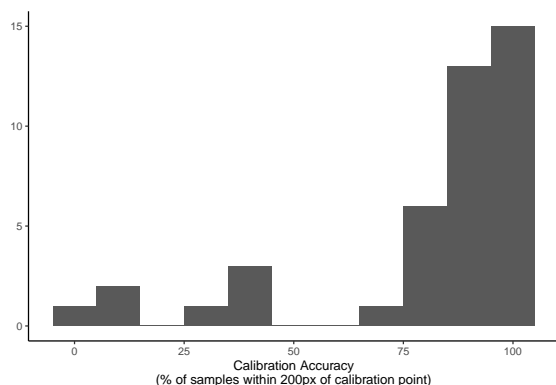


*Figure 24*. Histogram of calibration success in Exp. 5. Where participants required more than one calibration (N=8), only the final calibration was considered.

933     To determine whether calibration accuracy influenced our key effects, we calculated

934  the percentage of samples during the task in which the participant was fixating the face

935  they ultimately chose. There was a significant correlation for both the attractiveness

936  judgments (r = 0.47 [0.04, 0.75], p = 0.03) and the roundness judgments (r = 0.60 [0.23,

937  0.82], p = 0). Inspection of Fig. 25 reveals that this correlation is due to a handful of

938  particpiants with calibration values below 50%.

939     Thus, we re-analyzed the data, removing the participants whose calibration accuracy

940  was not greater than 50%. This slightly improved the fits of the sigmoids (Attractiveness:

941  $R^2 = 0.79$; Roundness: $R^2 = 0.60$). However, the difference between sigmoids remained

942  non-significant using the Kolmogorov-Smirnov test (D = 0.22, p = 0.36). Descriptively, the

943  results do not look substantially different (Fig. 26).

944     **Effects of ROIs.**  In the original experiment, eye gazes that did not directly fixate

945  one or other of the faces were excluded. In this section we explore an alternative coding of

946  the eye movement data by coding simply left half vs. right half of the screen. The coarser

947  coding may be more appropriate for webcam-based eyetracking.

948     Only a small percentage of samples (7.00%) involved looks to anything other than
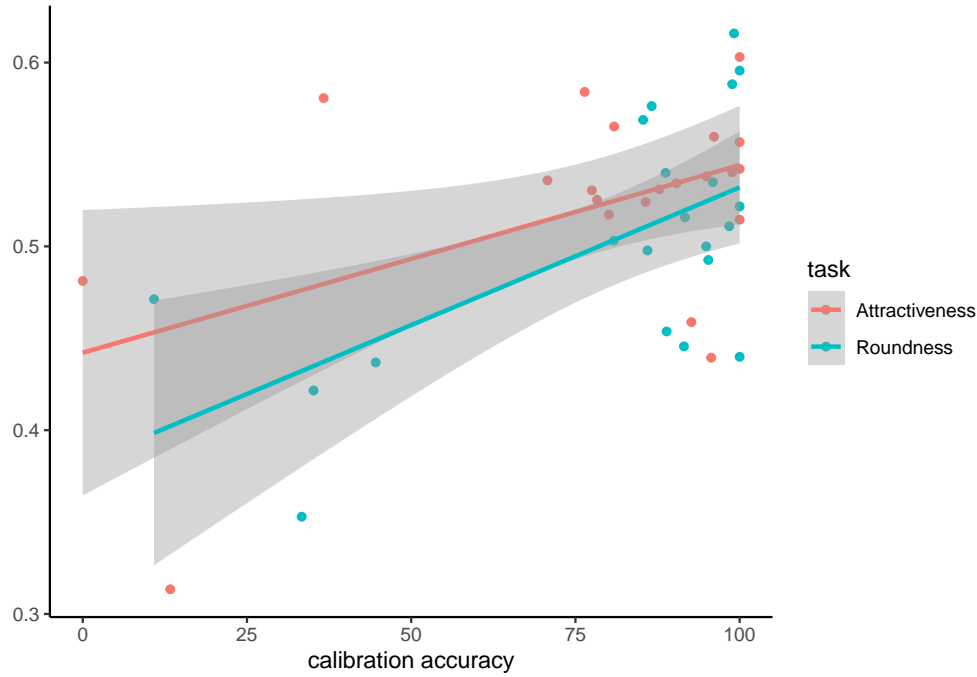
*Figure 25*. Correlation between calibration accuracy (x-axis) and percentage of samples fixating target (y-axis) in Exp. 5.
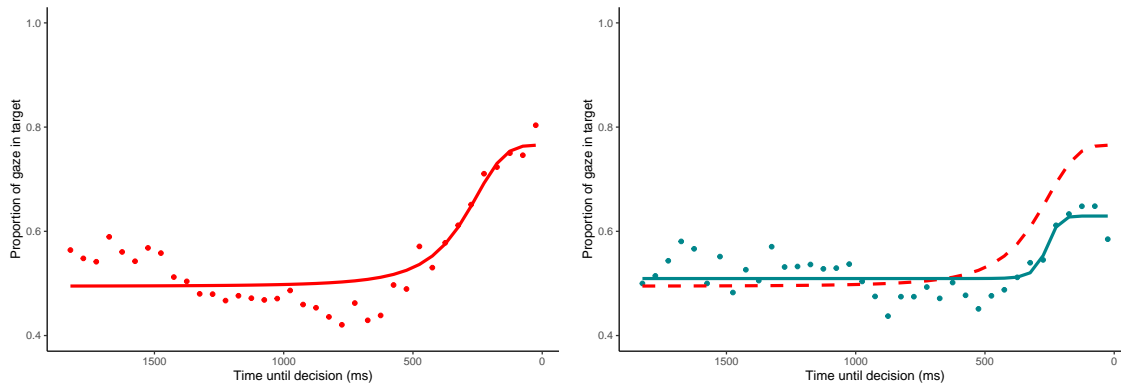


*Figure 26*. Revised results for Exp. 5 after removing low-calibration accuracy participants. *Left*: Eyegaze during attractiveness judgments, along with the best-fitting sigmoid. *Right*: Eyegze during roundness judgments, along with best-fitting sigmoid (best-fitting sigmoid for attractiveness is re-plotted with a dashed line for comparison).

949 one of the two faces. Thus, not surprisingly, the correlation between percentage of time

950 spent fixating the to-be-chosen face using the ROI method and the halves method was near

951 ceiling (r = 0.97 [0.97, 0.98], p = 0). Since the choice of method had almost no effect on

952 whether participants were coded as fixating one face or the other, we did not further

953 investigate the effect of method choice on the analytic results.

## Discussion

955 Qualitatively, the results are similar to those of Shimojo et al., such that participants

956 look more at the option that they ultimately choose. This gaze bias appears to be stronger

957 for decisions about face attractiveness than shape, though this is not supported by the

958 statistical analysis approach used in the original paper. The gaze patterns remained

959 consistent for participants with better calibration accuracy.

## General Discussion

961 We conducted five attempted replication studies using different experimental

962 paradigms from across the cognitive sciences. All were successfully implemented in `jsPsych`

963 using the `webgazer` plugin, but replication success was mixed. Experiment 1 had the

964 smallest ROIs due to the use of an integrated visual scene with five to six ROIs of varying

965 size per scene, as opposed to ROIs corresponding to display halves or quadrants. Both

966 attempts to replicate Altmann and Kamide (1999) were unsuccessful, despite the success of

967 previous in-lab replications using infrared eye-tracking (e.g. James et al., 2023). A previous

968 conceptual replication of this paradigm using webcam-based eye-tracking (Prystauka et al.,

969 2023) was successful but used a four-quadrant visual world paradigm, rather than the

970 "naturalistic" scenes used in the original study and in the current replication attempts. It

971 is worth noting that removing variability related to participant environments (by

972 conducting the webcam-tracking study in the lab) did not appear to improve the

973 sensitivity of the paradigm. The primary limitation is likely to be the size of the ROIs.

974   Experiment 2 used the four quadrants of the participant's screen as ROIs. As in

975   Johansson and Johansson (2014) and Spivey and Geng (2001), participants spontaneously

976   looked to blank ROIs which previously contained to-be-remembered pictures. These results

977   appeared to be robust to calibration quality. An additional manipulation, instructing

978   participants to keep gaze fixed on a central point, was not successful. One possibility is

979   that participants are less motivated to follow such instructions when an experimenter is not

980   present in the same room with them. It may be possible to improve performance by

981   emphasizing that this is an important aspect of the experiment or by providing additional

982   training/practice in keeping the eyes still on one particular point.

983   Experiment 3 used two large ROIs (halves of the display in one analysis) and

984   successfully replicated the novelty preference in terms of gaze duration shown in Manns et

985   al. (2000). However, the subtler relationship between gaze duration and recognition

986   memory on Day 2 was not replicated, despite the fact that participants were able to

987   discriminate pictures they had seen from those they hadn't seen during that delayed test.

988   Calibration quality did not appear to impact this relationship. More work is needed to

989   understand whether delay manipulations can be practically combined with webcam

990   eye-tracking.

991   Experiment 4 used the four quadrants of the participant's screen as ROIs. As in

992   Ryskin et al. (2017), listeners used knowledge of the co-occurrence statistics of verbs and

993   syntactic structures to resolve ambiguous linguistic input ("Rub the frog with the mitten").

994   Across multiple time windows, participants looked more at potential instruments (mitten),

995   when the verb (rub) was one that was more likely to be followed by a prepositional phrase

996   describing an instrument with which to perform the action , as opposed to describing the

997   recipient of the action (frog). Despite the qualitative replication of past findings, the

998   overall rates of looks to various objects were much lower than in an in-lab study using

999   infrared eye-tracking. This reduction may be related to measurement quality: effect sizes

1000  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 et al. (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**

1023

1024  Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting

1025      the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

1026      https://doi.org/10.1016/S0010-0277(99)00059-1

1027  Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020).

1028      Gorilla in our midst: An online behavioral experiment builder. *Behavior Research*

1029      *Methods*, *52*, 388–407.

1030  Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*.

1031      Retrieved from https://github.com/crsh/papaja

1032  Bánki, A., Eccher, M. de, Falschlehner, L., Hoehl, S., & Markova, G. (2022). Comparing

1033      online webcam-and laboratory-based eye-tracking for the assessment of infants'

1034      audio-visual synchrony perception. *Frontiers in Psychology*, *12*, 733933.

1035  Barth, M. (2022). *tinylabels: Lightweight variable labels*. Retrieved from

1036      https://cran.r-project.org/package=tinylabels

1037  Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects

1038      models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

1039      https://doi.org/10.18637/jss.v067.i01

1040  Bates, D., & Maechler, M. (2021). *Matrix: Sparse and dense matrix classes and methods*.

1041      Retrieved from https://CRAN.R-project.org/package=Matrix

1042  Bolker, B., & Robinson, D. (2020). *Broom.mixed: Tidying methods for mixed models*.

1043      Retrieved from https://CRAN.R-project.org/package=broom.mixed

1044  Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., . . . Borges, B. (2021).

1045      *Shiny: Web application framework for r*. Retrieved from

1046      https://CRAN.R-project.org/package=shiny

1047  de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments

1048      in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

1049      https://doi.org/10.3758/s13428-014-0458-y

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.

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

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.

Kukona, A., Cho, P. W., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 326.

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

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

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.

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.

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*

1104    *Experimental Psychology: General, 152*(5), 1245.

1105    Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive

1106    science: A first look. *Behavior Research Methods, 50*, 451–465.

1107    Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and

1108    influences preference. *Nature Neuroscience, 6*(12), 1317–1322.

1109    https://doi.org/10.1038/nn1150

1110    Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication

1111    results. *Psychological Science, 26*(5), 559–569.

1112    https://doi.org/10.1177/0956797614567341

1113    Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2021). *Afex:*

1114    *Analysis of factorial experiments.* Retrieved from

1115    https://CRAN.R-project.org/package=afex

1116    Slim, M. S., & Hartsuiker, R. J. (2022). Moving visual world experiments online? A

1117    web-based replication of dijkgraaf, hartsuiker, and duyck (2017) using PCIbex and

1118    WebGazer. js. *Behavior Research Methods*, 1–19.

1119    Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions:

1120    The role of lexical-biases and referential scenes in child and adult sentence processing.

1121    *Cognitive Psychology, 49*(3), 238–299. https://doi.org/10.1016/j.cogpsych.2004.03.001

1122    Spivey, M. J., & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and

1123    memory: Eye movements to absent objects. *Psychological Research, 65*(4), 235–241.

1124    https://doi.org/10.1007/s004260100059

1125    Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A.,

1126    et al.others. (2024). Validation of an open source, remote web-based eye-tracking

1127    method (WebGazer) for research in early childhood. *Infancy, 29*(1), 31–55.

1128    Sun, C., & Breheny, R. (2020). Another look at the online processing of scalar inferences:

1129    An investigation of conflicting findings from visual-world eye-tracking studies.

1130    *Language, Cognition and Neuroscience, 35*(8), 949–979.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

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.

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