What paradigms can webcam eye-tracking be used for? Attempted replications of 5 "classic" cognitive science experiments

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

[1] Cognitive Science Department, Vassar College
[2] Department of Cognitive & Information Science, University of California, Merced
[3] Psychology Department, Macalester College
[4] Department of Psychology & Neuroscience, Boston College

Author Note

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

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

Abstract

ADD LATER

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

Word count: X

What paradigms can webcam eye-tracking be used for? Attempted replications of 5 "classic" cognitive science experiments

Intro stuff:

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

Present work

In order to validate online eyetracking measures, we set out to reproduce five previously published studies representing a variety of questions, topics, and paradigms. The goal was to examine the strengths and weaknesses of webcam eye-tracking for common paradigms in cognitive science. Ideally, we would only attempt to replicate studies where the original measurements have small error bars and are known to replicate; otherwise, it can be difficult to distinguish a failure of the method (online eyetracking does not work) from a failure of the original study to replicate.

In practice, replications (successful or otherwise) have only been reported for a small number of studies, so we ultimately included some studies with unknown replicability. We addressed this in several ways. First, replicating five very different studies from different research traditions decreases our reliance on any one study. Second, we include several "sanity check'' analyses, such as the correlation between calibration accuracy and effect size. (If the effect is real but there is noise from low-accuracy eyetracking, this correlation should be substantial.) Third, for XXX of the studies, we had comparison data collected in-lab either using jsPsych or a more traditional eyetracker technology, allowing us to direct assess the impact of differences in subject population and equipment.

Table 1

Studies selected for replication attempts

Citation	Topic Area	Paradigm	Citations (June 2022, Google Scholar)
Altmann & Kamide, 1999	Psycholinguistics	Natural Scenes	1,840.00
Johansson & Johansson, 2013	Memory	Four Quadrants	190.00
Manns, Stark, & Squire, 2000	Memory	Two Halves	127.00
Snedeker & Trueswell, 2004	Psycholinguistics	Four Quadrants	448.00
Shimojo et al., 2003	Decision Making	Two Halves	964.00

Selection of Studies

We chose five high-impact eyetracking studies involving adult subjects. (Given the additional difficulties of recruiting and retaining child participants, we excluded developmental studies.) 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 "natural" 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 **??**tab:studies-table) 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 order to have adequate statistical power and precision, we aimed for 2.5x the sample size of the original experiment, following the heuristic of Simonsohn (Simonsohn, 2015). In study 5, the original sample size was so small that we opted to collect 5x the number of participants to increase precision. Because of budget and time constraints we were unable to replace the data for subjects who were excluded or whose data was missing due to technical failures.

Equipment

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

Eye-tracking Calibration and Validation

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

During the calibration, dots appeared on the screen one at a time in different locations, and the participants had to fixate them and click on each one. Once they clicked on a dot, it would disappear and a new one would appear in a different location on the screen. The locations of calibration dots were specific to each experiment (details below) and appeared in the areas of the screen where the visual stimuli would appear during the main task in order to ensure that eye movements were accurately recorded in the relevant regions of interest. After the calibration was completed, the validation began. Participants were asked to go through the same steps as the calibration, except that they only fixated the dots as they appeared in different locations on the screen. If accuracy on the validation was too low (fewer than 50% of looks landed within a 200 px radius of the validation points), participants were given an opportunity to re-start the calibration and validation steps. If the second attempt also lead to low validation accuracy, participants were informed that they could not participate in the study.

Data pre-processing

We used R (Version 4.2.1; R Core Team, 2021) and the R-packages *afex* (Version 1.1.1; Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2021), *broom.mixed* (Version 0.2.9.4; Bolker & Robinson, 2020), *dplyr* (Version 1.0.10; Wickham, François, Henry, & Müller, 2021), *forcats* (Version 0.5.2; Wickham, 2021a), *ggplot2* (Version 3.3.6; Wickham, 2016), *jsonlite* (Version 1.8.0; Ooms, 2014), *lme4* (Version 1.1.30; Bates, Mächler, Bolker, & Walker, 2015), *lmerTest* (Version 3.1.3; Kuznetsova, Brockhoff, & Christensen, 2017), *Matrix* (Version 1.5.1; Bates & Maechler, 2021), *papaja* (Version 0.1.1; Aust & Barth, 2020), *readr* (Version 2.1.2; Wickham & Hester, 2020), *shiny* (Chang et al., 2021), *stringr*

(Version 1.4.1; Wickham, 2019), *tidyr* (Version 1.2.1; Wickham, 2021b), and *tinylabels* (Version 0.2.3; 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, Spivey-Knowlton, Eberhard, & Sedivy, 1995) to show that meanings of verbs rapidly constrain the set of potential subsequent referents in sentence processing. For example, when looking at the display in Figure 2 and listening to a sentence like "The boy will eat the…," participants are more likely to look at the cake than when they hear "The boy will move the…," in which case they tend to look at the train, presumably because cakes are edible and trains are not. Semantic information available at the verb is used to anticipate upcoming linguistic input.

Methods

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

Participants. 60 participants were paid $2.60 for their participation. Our sample size of participants was determined by the total run time of our experiment, ~10 minutes, and the allotted funding from the Vassar College Cognitive Science Department. From this information, we calculated a reasonable number of participants we could afford to compensate on Prolific. Note that the sample size of the original study was 24. For unknown reasons, 2 of the subjects' results were not recorded, so in the analysis, we worked with data collected from 58 participants.

Procedure. The task began with a 9-point eye-tracker calibration and validation (Figure **??**). During the experiment, the participants were simultaneously presented with a visual image and a corresponding audio recording of a spoken sentence. Participants had to input a keyboard response indicating “yes” or “no” as to whether the sentence they heard was feasible given the visual image. There were two practice trials to ensure that participants understood the instructions before they undertook the main portion of the experiment. Participants’ reaction times, keyboard responses, and looks to objects in the scene were recorded for each trial.

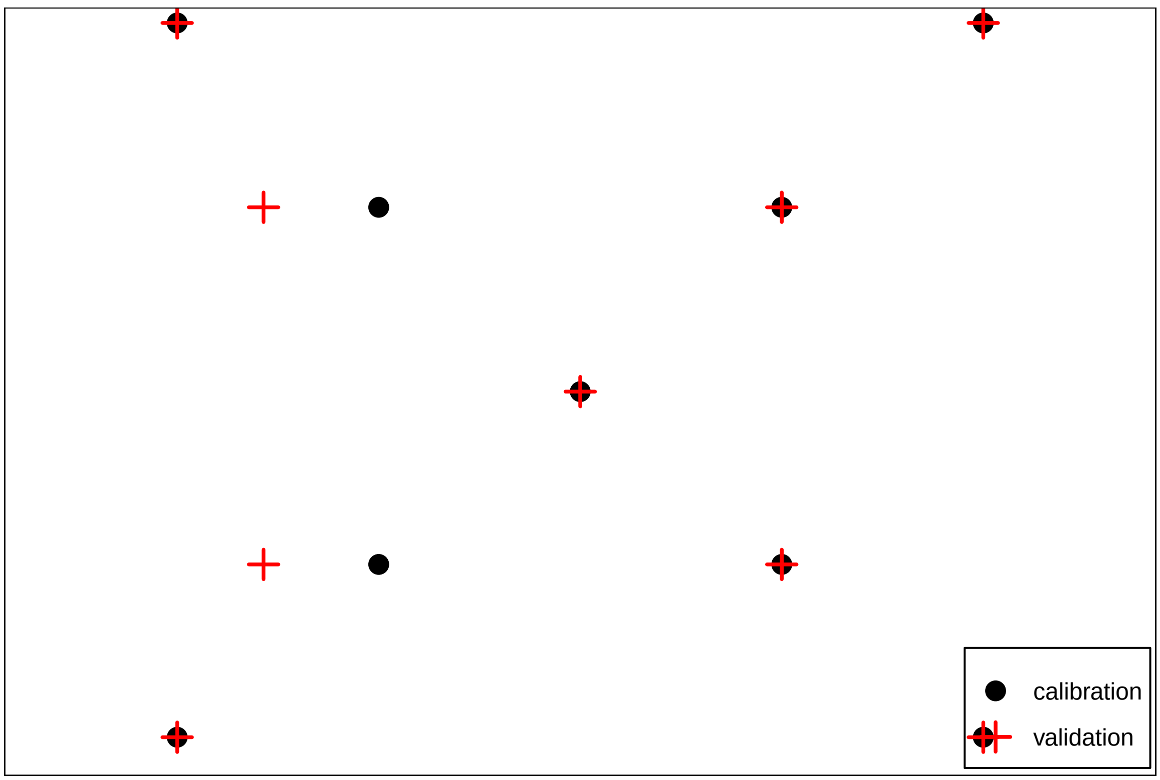


Figure 1. Calibration and validation point locations for Experiment 1. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.

Materials & Design. The visual stimuli were created through Canva and depicted an agent accompanied by four to five objects in the scene (see Figure 2). On critical trials, participants heard one of two sentences associated with the scene. In the restrictive condition, the sentence (e.g., “The boy will eat the cake”) contained a verb (e.g., “eat”) which restricts the set of possible subsequent referents (e.g., to edible things). Only the target object (e.g., the cake) was semantically consistent with the verb’s meaning. In the non-restrictive condition, the sentence (e.g., “The boy will move the cake”) contained a verb (e.g., “move”) which does not restrict the set of possible subsequent referents. The target object (e.g., the cake) as well as the distractor objects (e.g., the train, the ball, etc.) were semantically consistent with the verb’s meaning. Both sentences were compatible with the scene, such that the correct keyboard response for the critical trials was “yes.” Filler trials consisted of scenes that looked similar to critical scenes but were paired with inappropriate sentences. The correct keyboard response for the filler trials was “no.”

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.

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.

Results

Replication.

- here we will describe the analyses that are as close as possible to the original paper with a minimal validation cutoff
- same analysis but with stricter validation cutoff

Comparison to in-lab data.

- here we will describe a direct comparison to data collected in the lab

Calibration.

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

Discussion

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.

Methods

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 **??**).

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 4). 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 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.

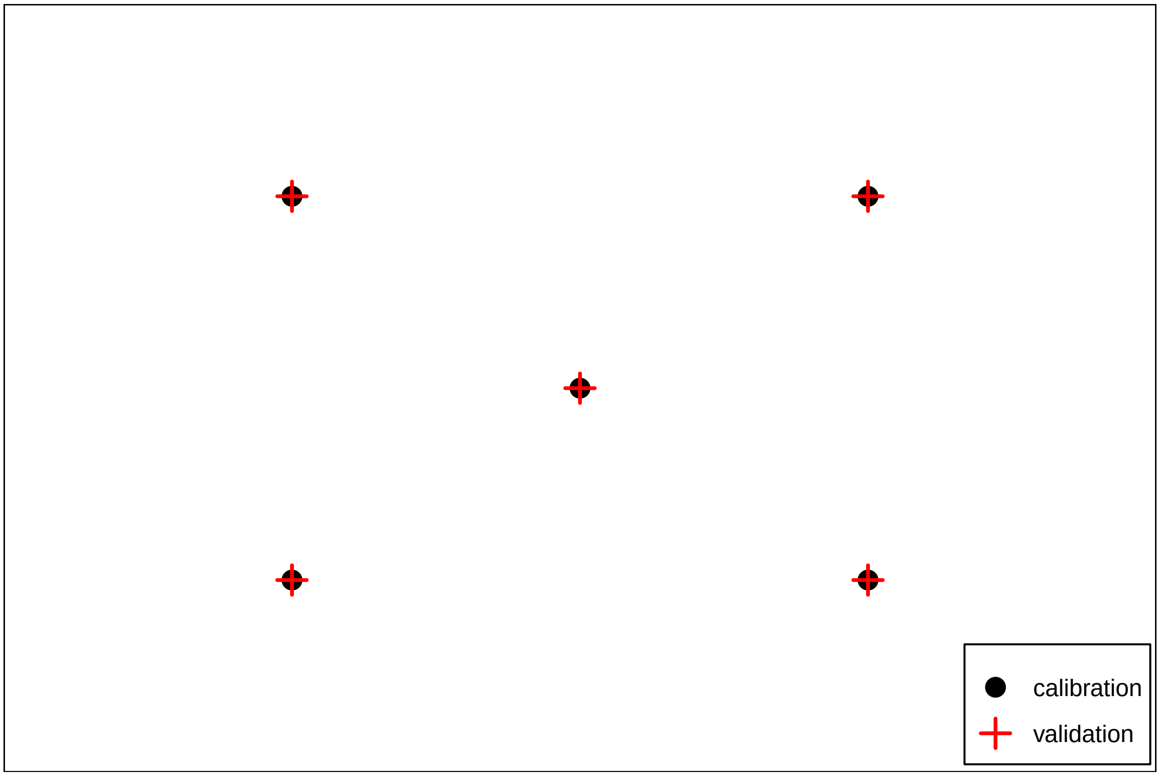


Figure 3. Calibration and validation point locations for Experiment 2. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.(In this experiment all the same locations were used for both calibration and validation.)

Figure 4. Example trial from Experiment 2.

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 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

Replication. **Eye-gaze**. Looks during the retrieval period were categorized as belonging to one of four quadrants based on the x,y coordinates. The critical quadrant was the one in which the to-be-retrieved object had been previously located during encoding. The other three quadrants were semi-randomly labeled "first", "second," third" (e.g., when the critical quadrant was in the top left, the “first” quadrant was the top right quadrant, but when the critical quadrant was in the top right, “first” corresponded to bottom right, etc.). In both the fixed- and free-viewing condition, participants directed a larger proportion of looks to the critical quadrant (see Figure 5). This bias appeared larger in the free-viewing condition, suggesting that the manipulation was (somewhat) effective.

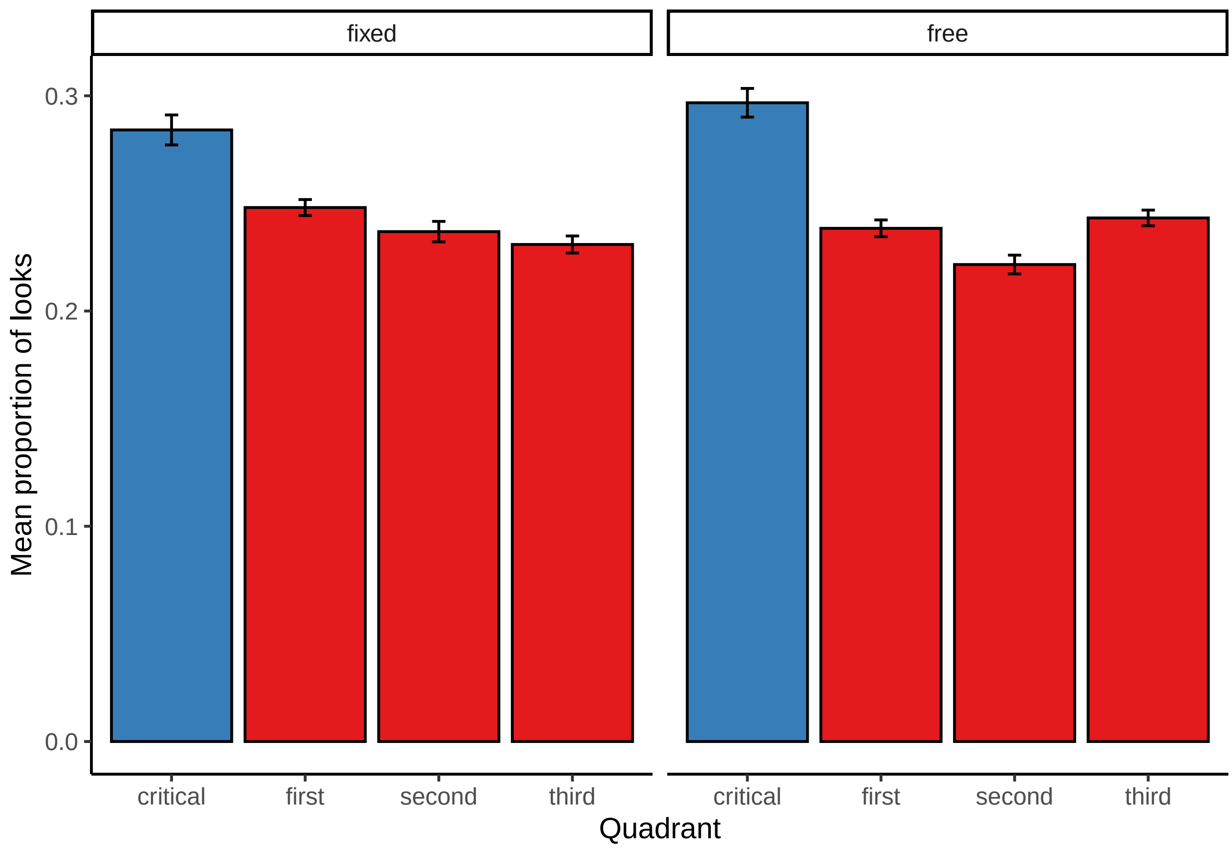


Figure 5. Proportion of eye-gaze to critical quadrant and other three quadrants during memory retrieval in a) fixed and b) free viewing conditions.

The proportions of looks across quadrants in the free-viewing condition were analyzed using a linear mixed-effects model with quadrant as the predictor (critical as the reference level). The model included random intercepts and slopes for participants[1]. Proportions of

[1] `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

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

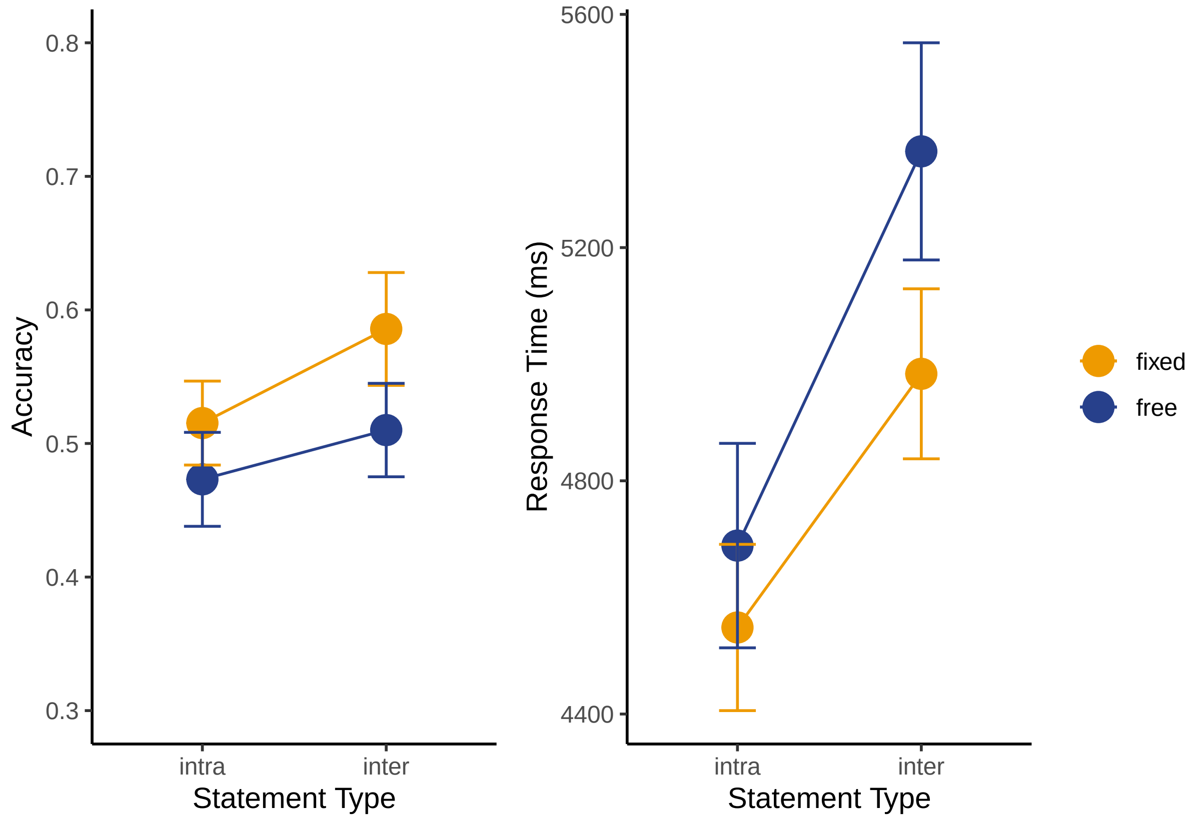


Figure 6. Accuracy and response times during memory retrieval.

Response Time and Accuracy. Participants' response times and accuracies on memory questions are summarized in Figure 6. Both dependent variables were analyzed with linear mixed-effects model with relation type (interobject = -0.5, intraobject=0.5) and viewing_condition (fixed = -0.5, free=0.5) and their interaction as the predictors. The model included random intercepts for participants[2]. Accuracy did not differ significantly

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

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

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.

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.

Methods

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 **??**).

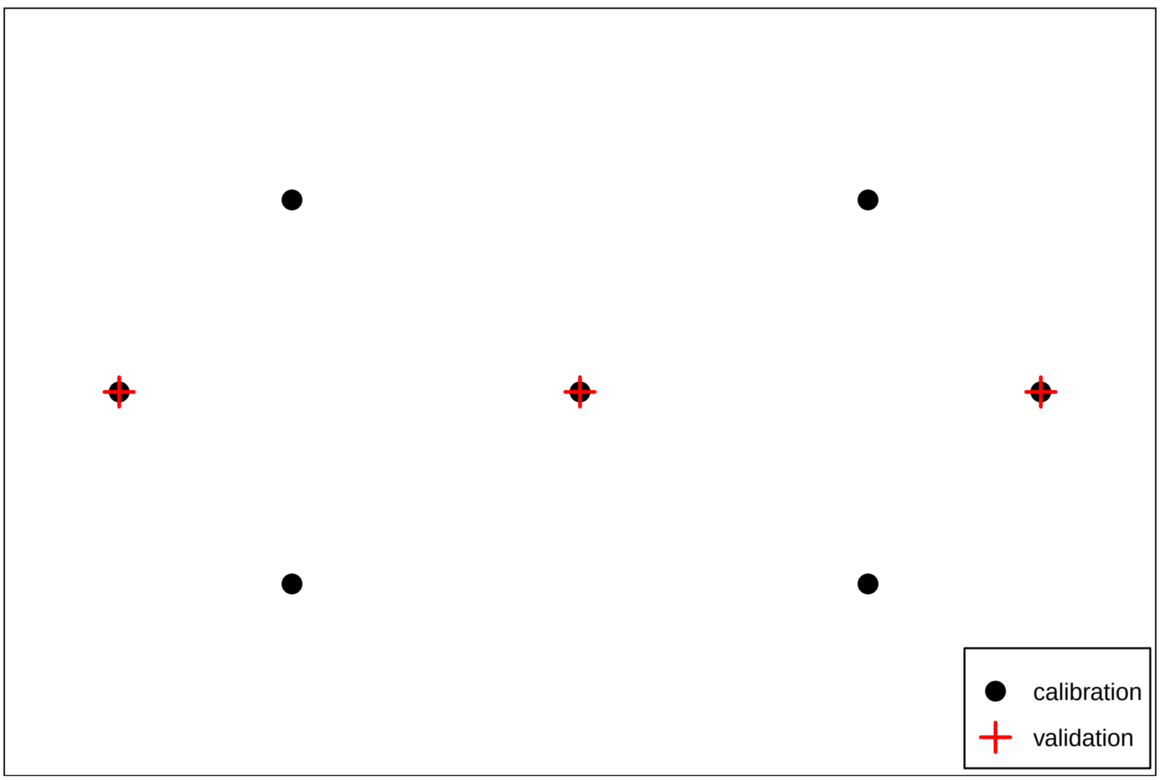


Figure 7. Calibration and validation point locations for Experiment 3. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.

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 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 8

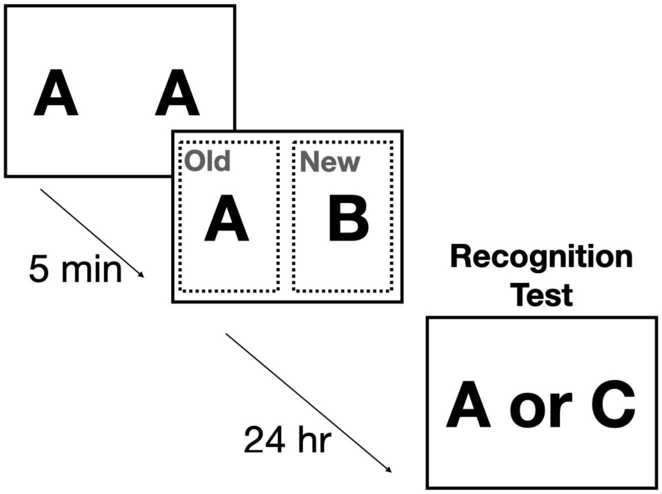


Figure 8. Schematic of the design of Experiment 3

Materials. Images were selected XXX...

There were two modifications we made to the methods of the original experiment. As we are 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 was presented on a single screen instead of two screens due to the constraints of the online experiment format.

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 novel objects.

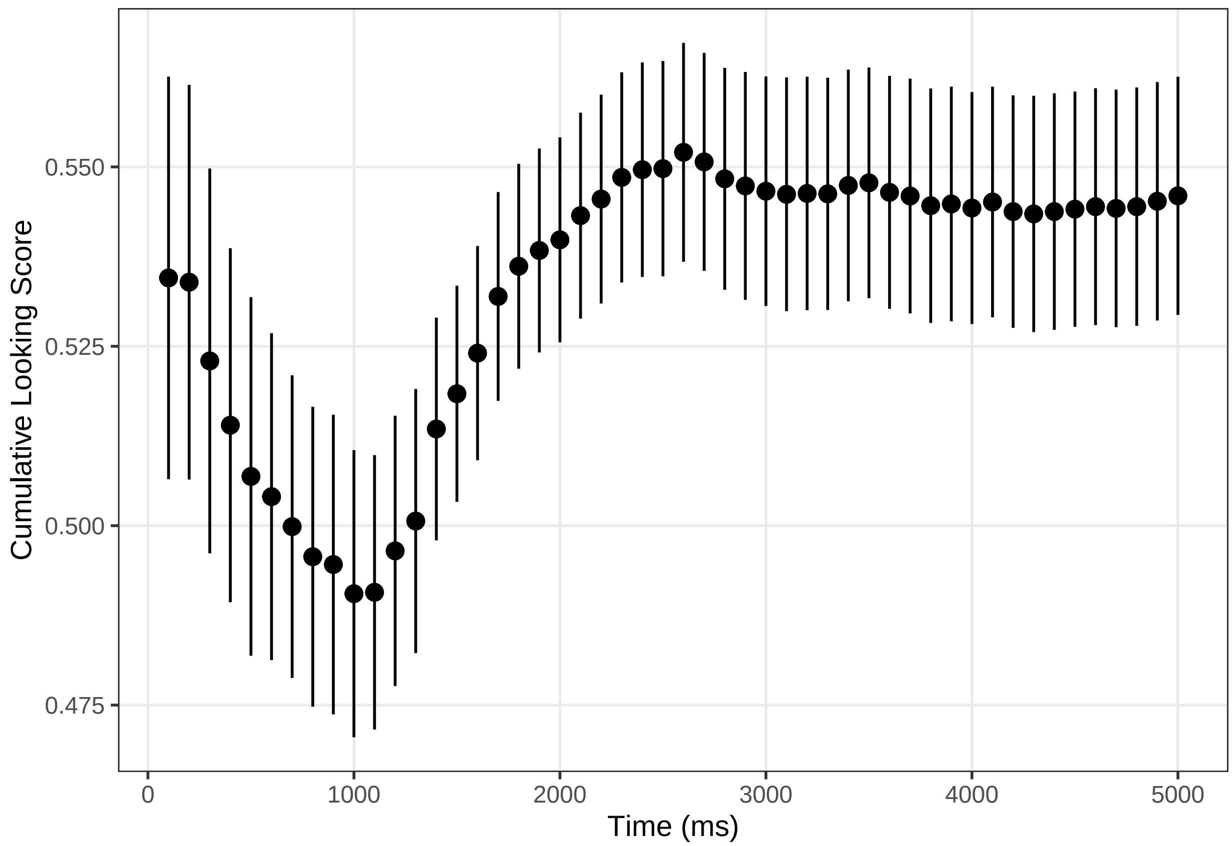


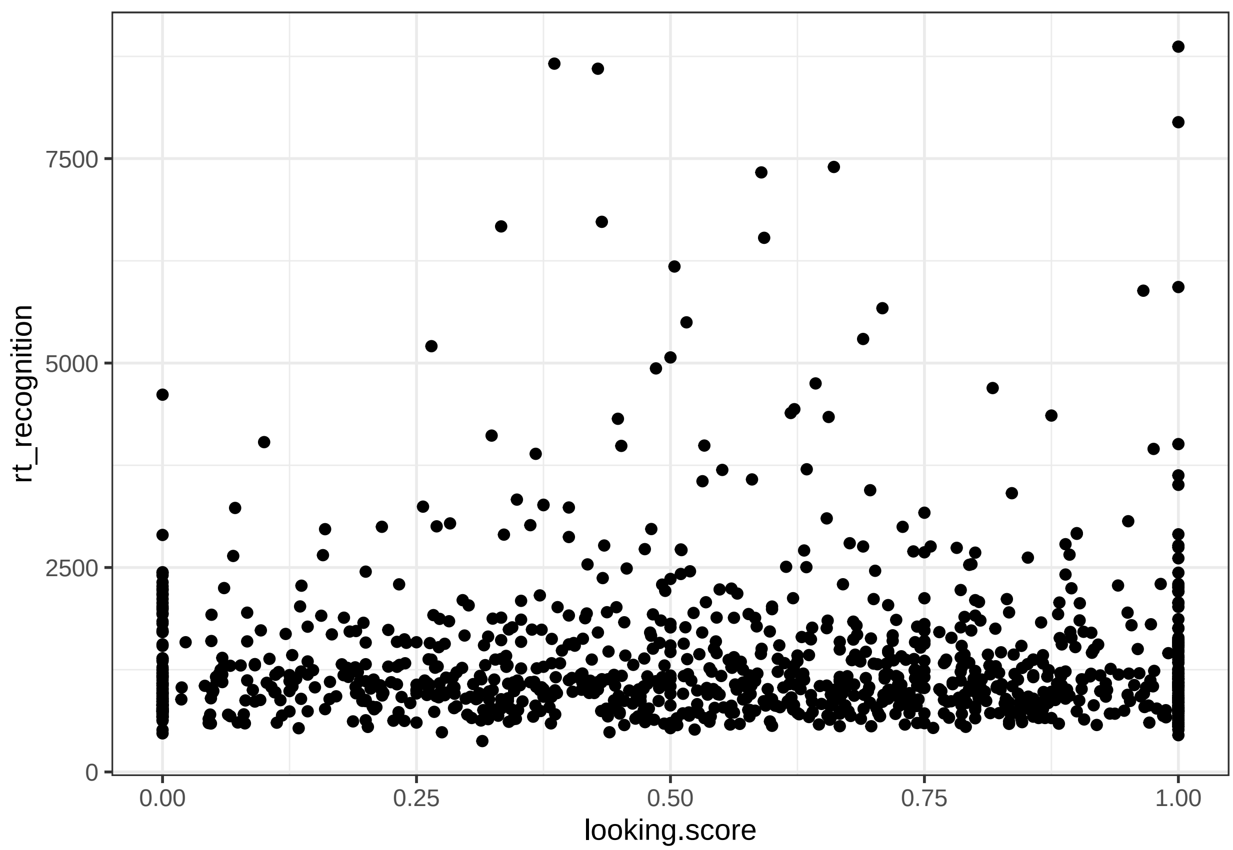
Figure 9. Cumulative looking score over the 5 second exposure during part 2 of day 1. Error bars represent +/- 1 SEM.

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

Participants correctly identified whether the image was familiar or unfamiliar 87.09% ($SD = 10.49$) of the time. After excluding the 3 participants who responded correctly to all images, the average confidence rating for correct responses (M = 3.51; SD = 0.41) was significantly higher than their average confidence ratings for incorrect responses (M = 2.55;

SD = 0.75), t(44) = -9.36, p = 0.00 . Among the same subset of participants, response times for correct responses (M = 1,443.49, SD = 413.94) were also significantly faster than for incorrect responses (M = 2,212.65, SD = 1,733.76), t(44) = 3.43 , p = 0.00.

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



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.

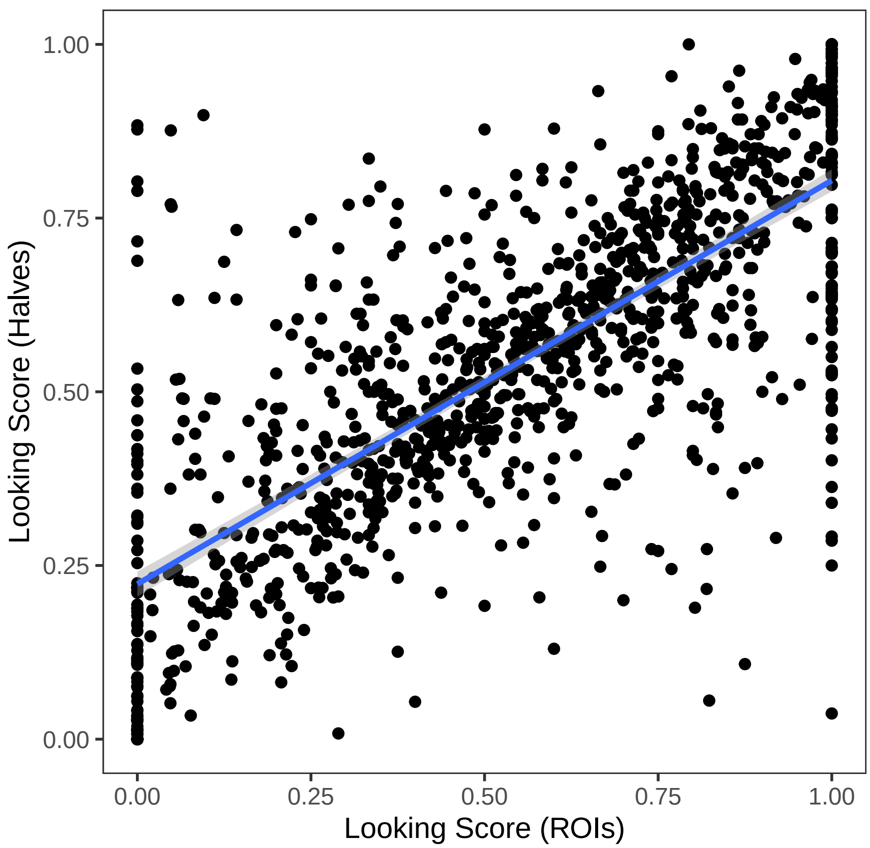


Figure 10. (#fig:E3-roi correlation of looking score)Correlation between looking scores calculated using ROIs and using screen halves.

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

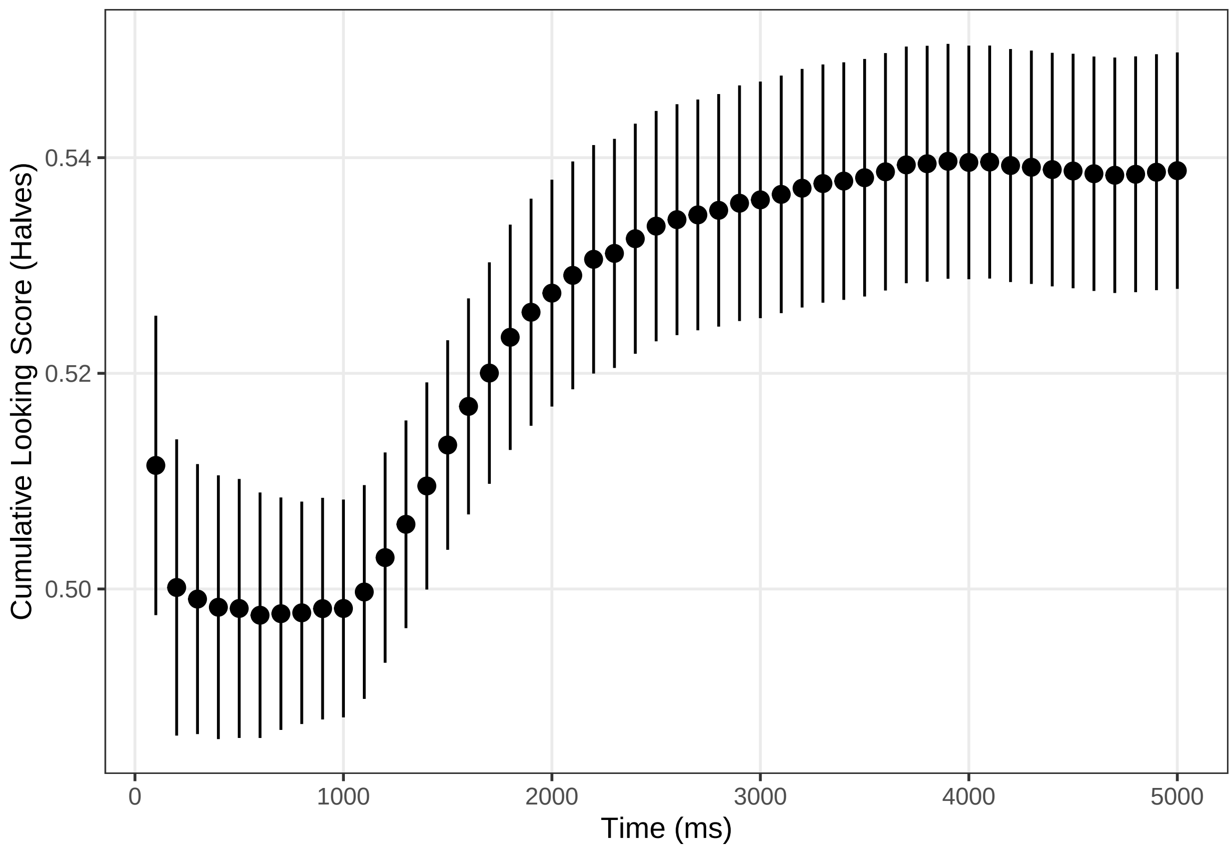
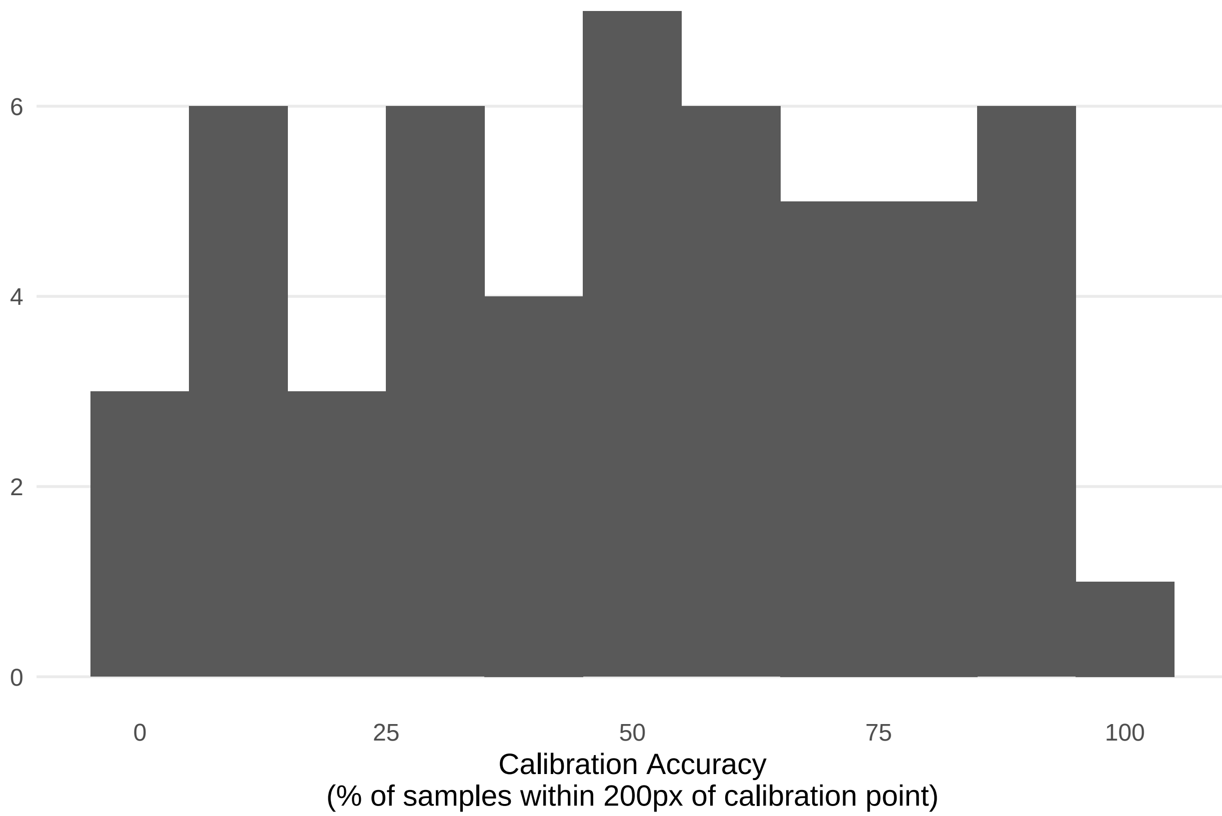


Figure 11. (#fig:E3-roi Plot of cumulative looking score)Cumulative looking score over the 5 second exposure during part 2 of day 1. Error bars represent +/- 1 SEM.

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

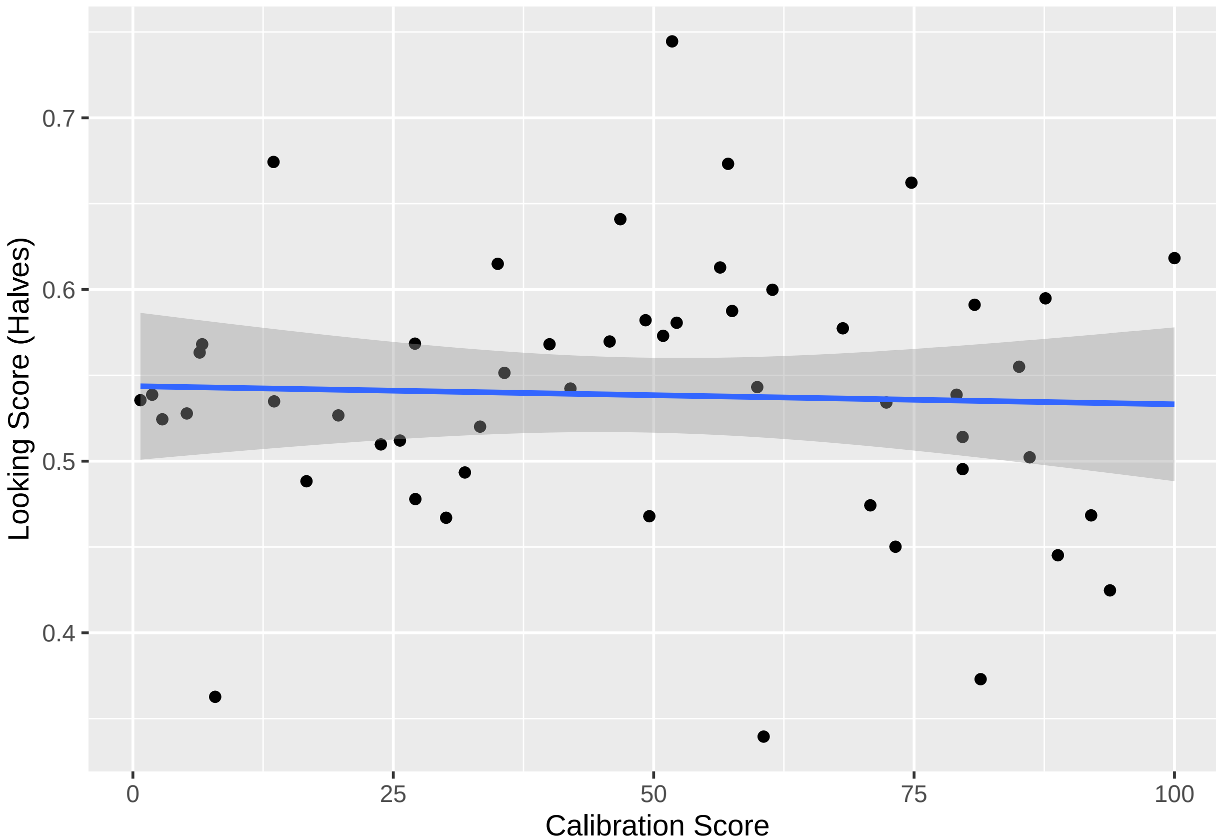
Calibration.

Calibration Accuracy.

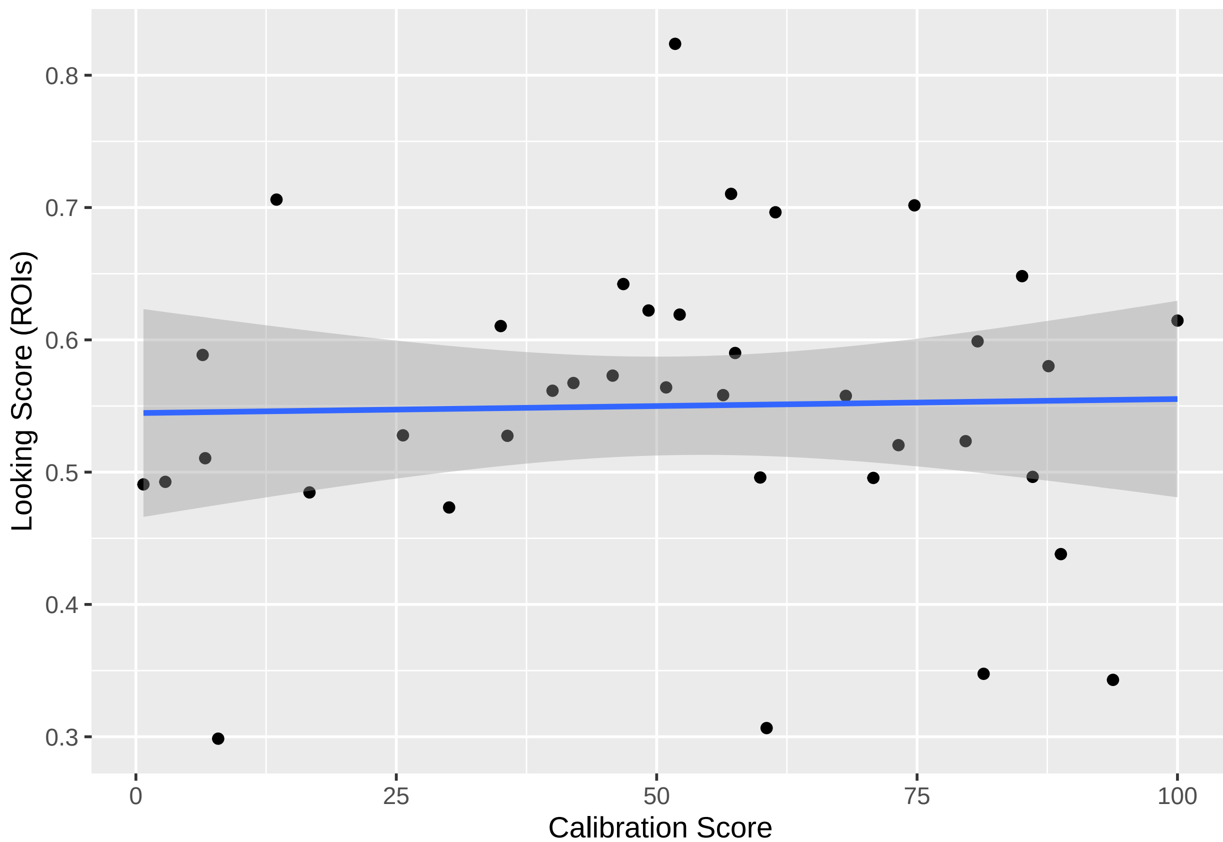


Correlation with Effects. To see if calibration success is correlated with the eye tracking effects, we calculated a calibration score for each participant. The calibration score was the average proportion of samples within XXX pixels of the validation points during the final validation phase before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

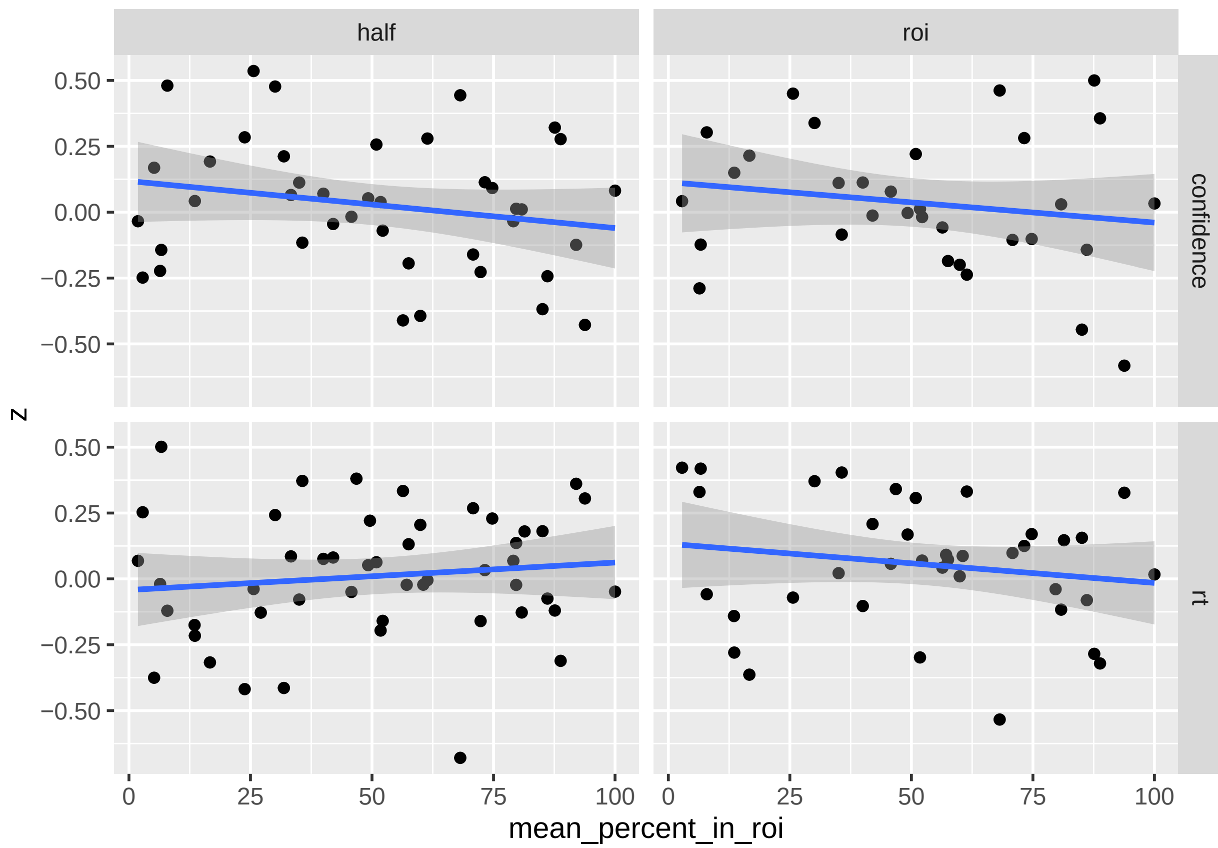
Calibration scores were not correlated with looking scores, regardless of which method was used to calculate looking scores.


0.7
0.6
0.5
0.4
Looking Score (Halves)
0
25
50
75
100
Calibration Score

411



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.



Discussion

Experiment 4

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

Methods

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

Participants. 57 participants were paid $2.50 for their participation. A sample size of 60 was initially chosen (but not reached in time) because we wanted to replicate the experiment with greater statistical power. Note that the original study had a sample size of 24.

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

During a trial, four pictures were displayed (target animal, target instrument, distractor animal, distractor instrument), one in each corner of the screen, and participants heard an audio prompt that contained instructions about the action they needed to act out (e.g., “Rub the butterfly with the crayon”; see Figure 13)[3]. Using their cursor, participants could act out the instructions by clicking on objects and moving them or motioning over the objects[4]. 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

[3] 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

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

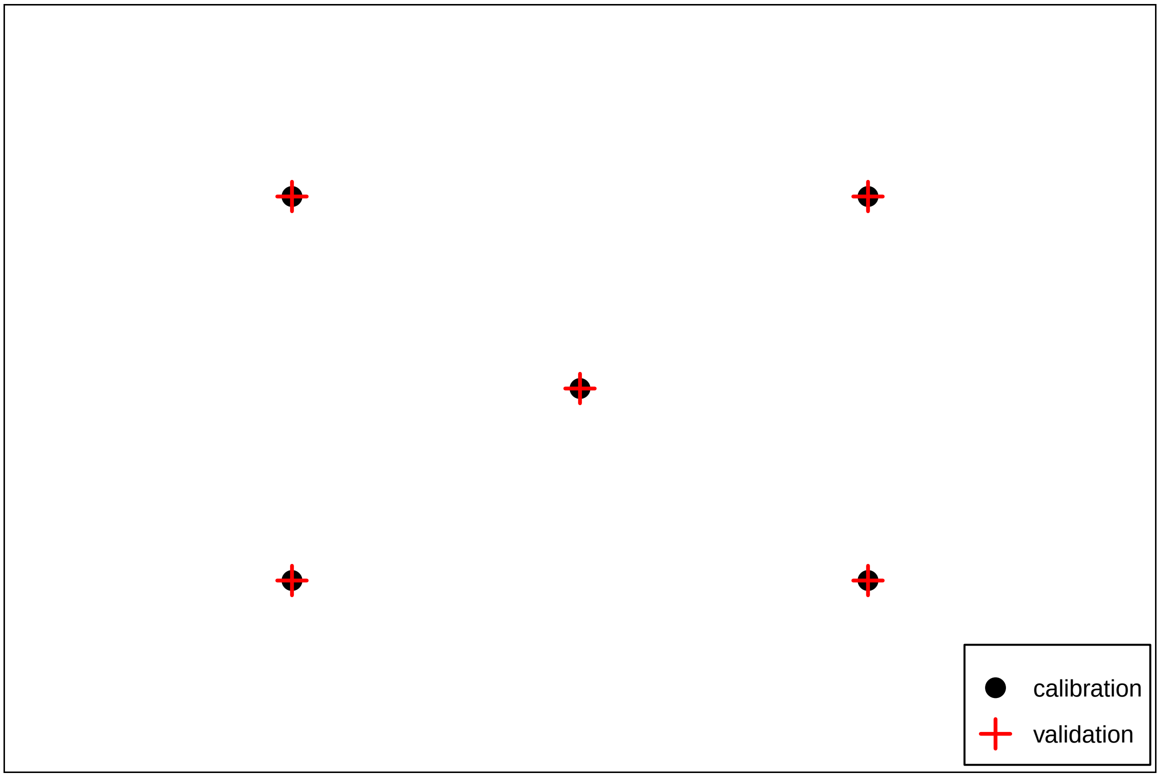


Figure 12. Calibration and validation point locations for Experiment 4. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.(In this experiment all the same locations were used for both calibration and validation.)

bias in the eye and mouse movements from the previous trial. The experiment only allowed the participants to move on to the next trial once the audio was completely done playing and the mouse had been moved over at least one object.

Materials. The images and audios presented to the participants were the same stimuli used in the original study (available here). The critical trials were divided into modifier-biased, instrument-biased, and equibiased conditions, and the filler trials did not contain ambiguous instructions. Two lists of critical trials were made with different verb and instrument combinations (e.g., "rub" could be paired with "panda" and "crayon" in one list and "panda" and "violin" in the second list). Within each list, the same verb was presented twice but each time with a different target instrument and animal. The lists were randomly assigned to the participants to make sure the effects were not caused by the

Figure 13. 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.

properties of the animal or instrument images used. The list of verbs used can be found in Appendix A of the original study.

Results

Replication. The location of initial mouse movements was used to assess whether the final interpretation of ambiguous sentences was biased by the verb. Figure 14 suggests that listeners were more likely to move their mouse first over the target instrument when the verb was equi-biased than when the verb was modifier-biased and even more so when the verb was instrument-biased. The opposite graded pattern can be observed for mouse movements over the target animal.

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

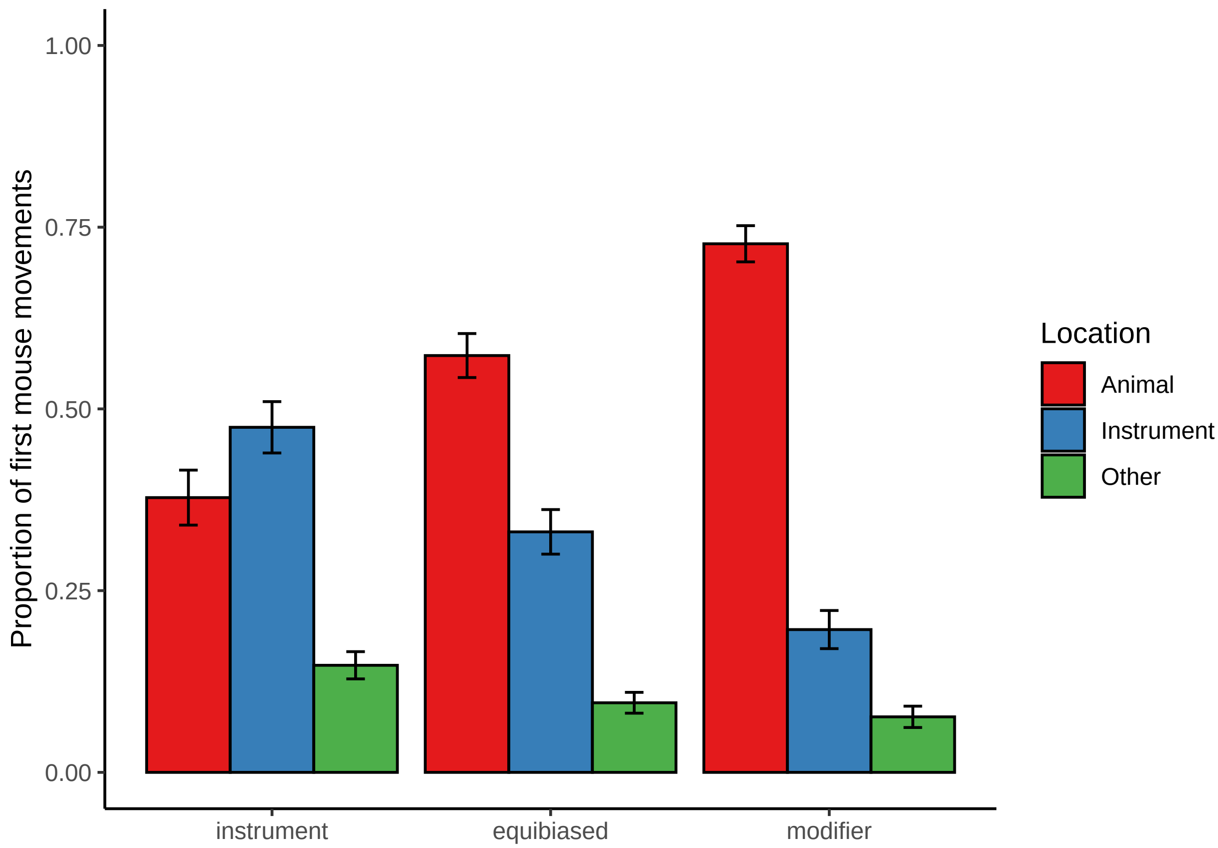


Figure 14. Proportion of first mouse movements by location and verb bias.

entered as varying intercepts with by-participant varying slopes for verb bias condition[5]. Participants were more likely to first move their mouse over target instruments in the instrument-biased condition relative to the equi-biased and modifier-biased condition (b = -1.50, SE = 0.25, $p < 0.01$). Further, participants were more likely to first move their mouse over target instruments in the equi-biased condition relative to the modifier-biased condition (b = -1.10, SE = 0.29, $p < 0.01$)

Gaze fixations were time-locked to the auditory stimulus on a trial by trial basis and categorized as being directed towards one of the four items in the display if the x, y coordinates fell within a rectangle containing the image. Figure 15 suggests that the

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

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

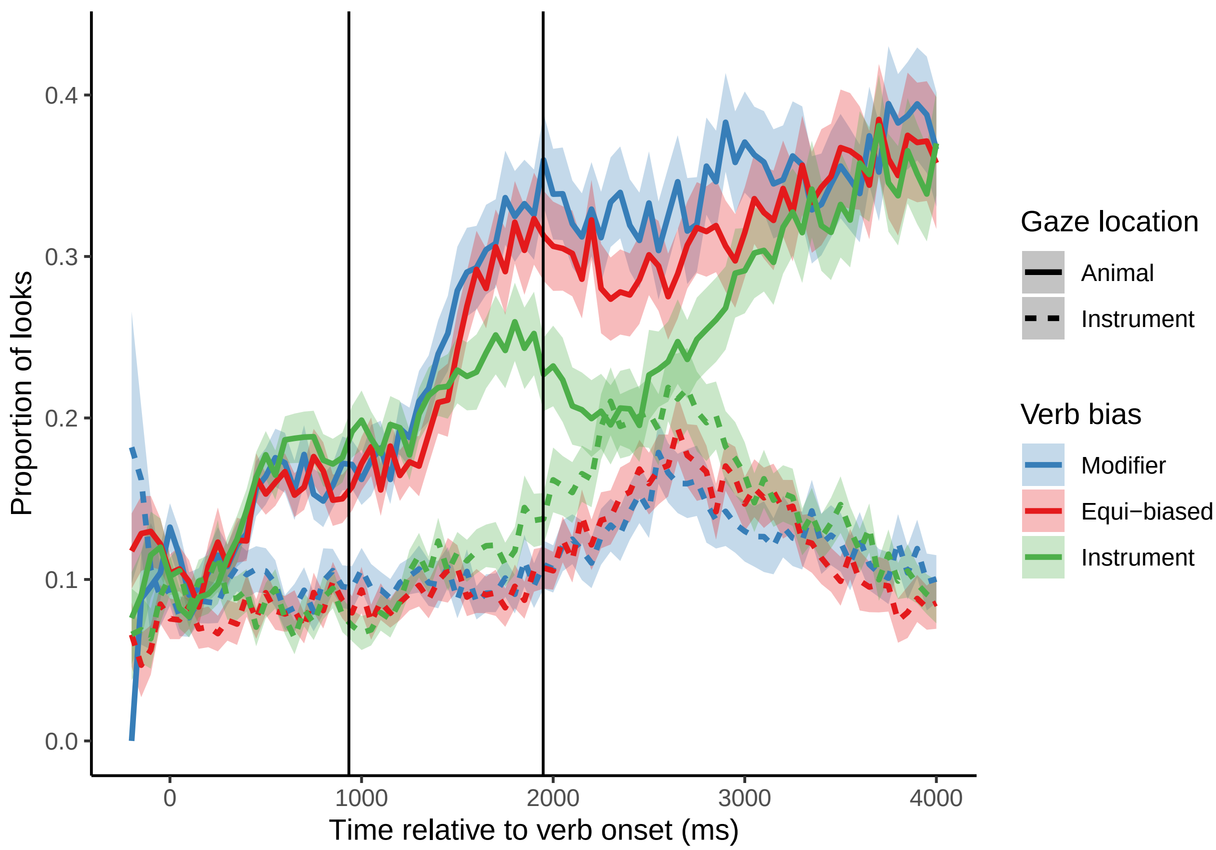


Figure 15. Timecourse of eye-gaze to target animal and target instrument by verb bias condition. Vertical lines indicate average onsets of animal and instrument offset by 200ms.

In order to assess how verb bias impacted sentence disambiguation as the sentence unfolded, the proportion of fixations was computed in three time windows: the verb-to-animal window (from verb onset + 200 ms to animal onset + 200 ms), the animal-to-instrument window (from animal onset + 200 ms to instrument onset + 200 ms), and the post-instrument window (from instrument onset + 200 ms to instrument onset + 1500ms + 200 ms). Mixed-effects linear regression models were used to predict the proportions of fixations to the target animal within each time window with the verb bias condition as an orthogonally contrast-coded (instrument vs. equi & modifier: inst = -2/3, equi = 1/3, mod = 1/3; equi vs. modifier: inst = 0, equi = -1/2, mod = 1/2) fixed effect. Participants and items were entered as varying intercepts[6]. In the *verb-to-noun* window, participants did not look more at the target animal in any of the verb bias conditions (Instrument vs. Equi and Modifier: b = -0.01, SE = 0.02, p = 0.59; Equi vs. Modifier: b = 0, SE = 0.02, p = 1). In the *noun-to-instrument* window, participants looked more at the target animal in the modifier-biased condition and equi-biased conditions relative to the instrument-biased condition (b = 0.03, SE = 0.01, $p < 0.01$) and in the modifier biased relative to the equi-biased condition (b = 0.02, SE = 0.01, $p < 0.05$). In the *post-instrument* window, participants looked more at the target animal in the modifier-biased condition and the equi-biased conditions relative to the instrument-biased condition (b = 0.08, SE = 0.02, $p < 0.01$) but not significantly so in the modifier biased condition relative to the equi-biased condition (b = 0.03, SE = 0.02, p = 0.15).

Comparison to in-lab data. The web version of the study qualitatively replicates the action and eye-tracking results of the original dataset (Ryskin et al., 2017). The mouse click results from both studies are summarized in Figure 16. The quantitative patterns of clicks were similar to those observed in the original dataset, though for Instrument-biased verbs, clicks were closer to evenly split between the animal and the instrument relative to the in-lab study where they were very clearly biased toward the instrument.

The eye-tracking results from both studies are summarized in Figure 17. For simplicity, and to reflect the dependent variable used in analyses, we average the proportion of fixations to the target animal within each time window. Though the qualitative patterns are replicated, proportions of fixations to the target animal were much

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

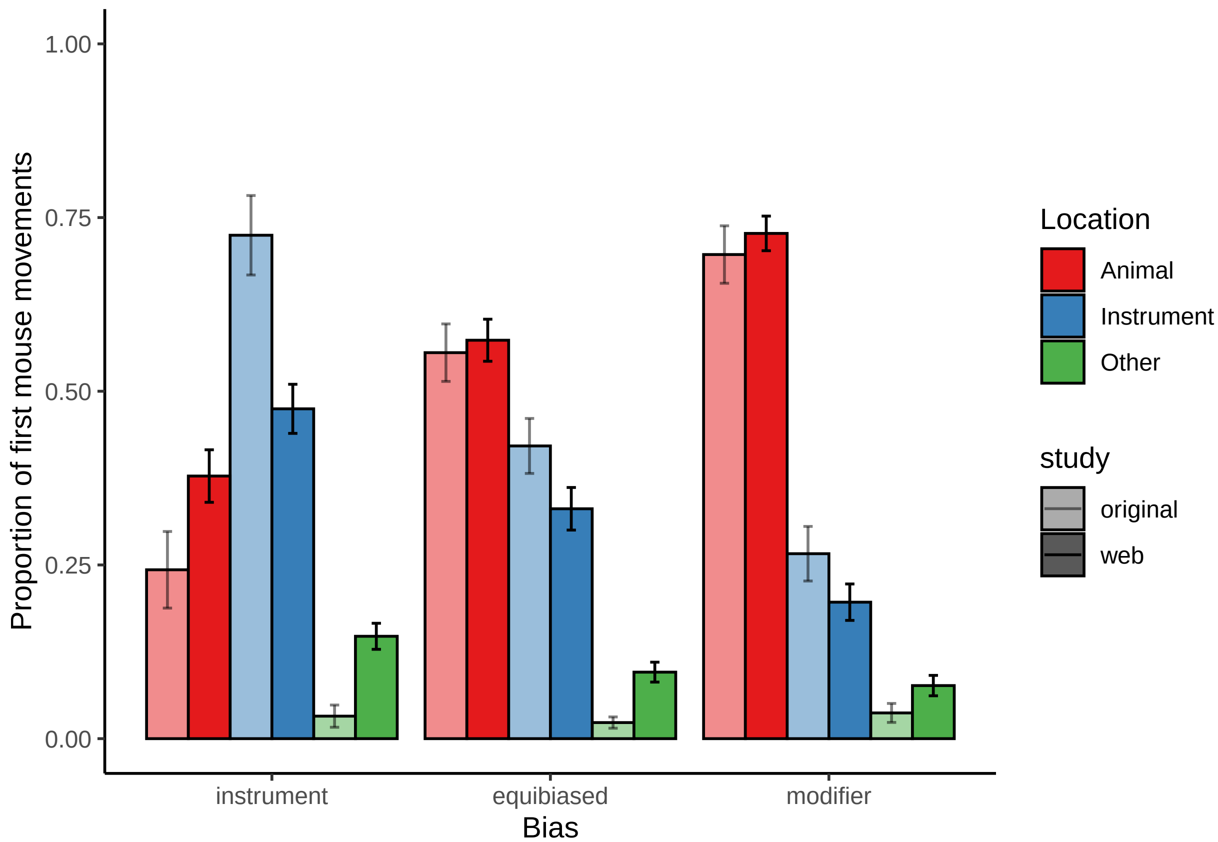


Figure 16. Proportion of first mouse movements by location and verb bias in the original dataset (Ryskin et al., 2017) and the current data collected online.

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

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

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

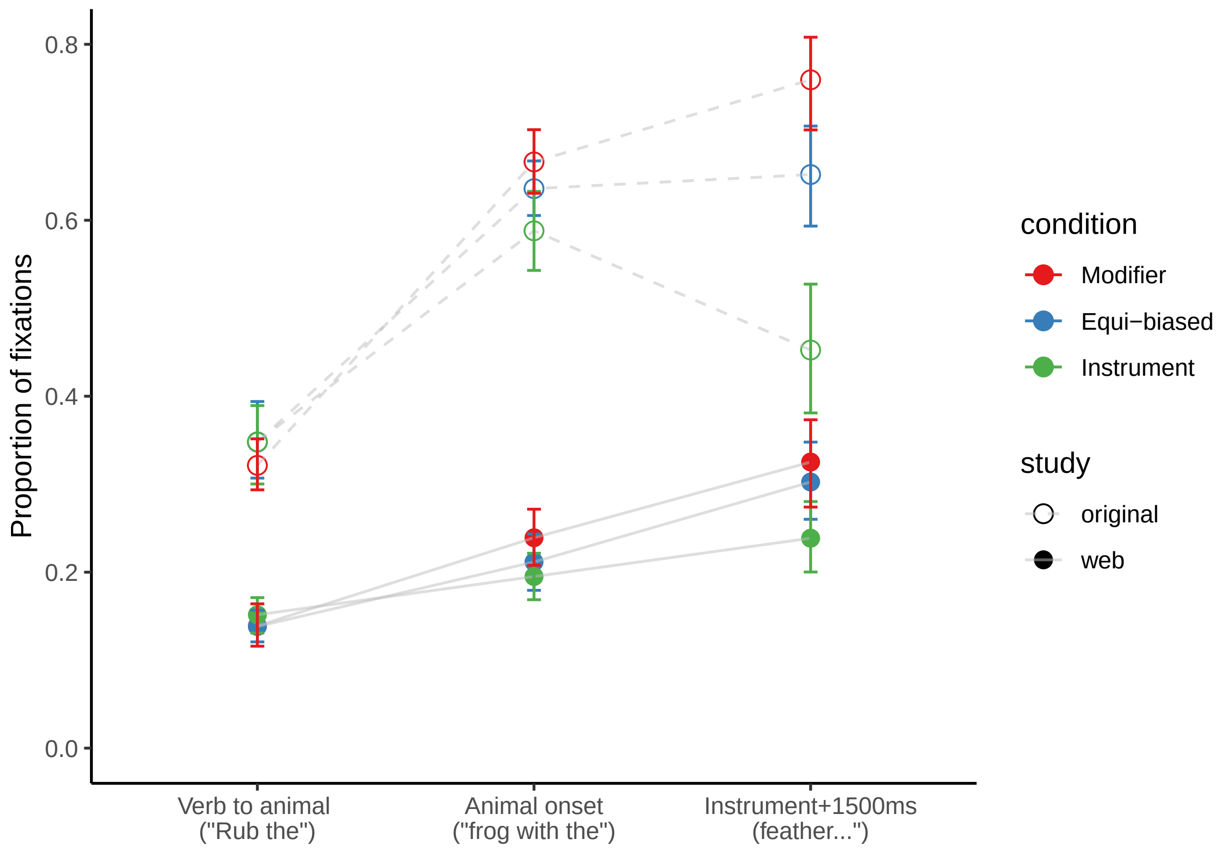


Figure 17. Proportion of target fixations by verb bias in the original dataset (Ryskin et al., 2017) and the current data collected online. Error bars reflect bootstrapped 95% CIs over subject means

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

Replicating the linear mixed-effects analysis (in the post-instrument onset time window only) on a subset of 19 participants with calibration quality >75% suggests that the effect of verb bias condition was larger in this subset than in the full dataset.

Participants looked more at the target animal in the modifier-biased condition and the equi-biased conditions relative to the instrument-biased condition ($b = 0.11$, $SE = 0.03$, $p < 0.001$) but not significantly so in the modifier biased condition relative to the equi-biased condition ($b = 0.05$, $SE = 0.03$, $p = 0.13$).

Effects of ROIs. Eye-tracking on the web differs critically from in-lab eye-tracking in that the size of the display differs across participants. Thus the size of the ROIs differs across participants. The current version of the web experiment used a bounding box around each image to determine the ROI. This approach is flexible and accomodates variability in image size, but may exclude looks that are directed at the image but fall outside of the image (due to participant or eye-tracker noise) as show in Figure 18a. Alternatively, The display can be split into 4 quadrants which jointly cover the entire screen (see Figure 18b).

Categorizing gaze location based on which of the four quadrants of the screen the coordinates fell in, increases the overall proportions of fixations (see Figure 19). In the *post-instrument* window, participants looked more at the target animal in the modifier-biased condition and the equi-biased conditions relative to the instrument-biased condition ($b = 0.08$, $SE = 0.02$, $p < 0.01$) and marginally so in the modifier biased condition relative to the equi-biased condition ($b = 0.04$, $SE = 0.02$, $p = 0.05$). Effect size estimates appeared somewhat larger and noise was somewhat reduced when using the quadrant categorization relative to the bounding box-based ROIs.

Discussion

Experiment 5

The fifth study was a replication attempt of Shimojo, Simion, Shimojo, and Scheier (2003), which found that human gaze is actively involved in preference formation. Separate sets of participants were shown pairs of human faces and asked either to choose which one they found more attractive or which they felt was rounder. Prior to making their explicit

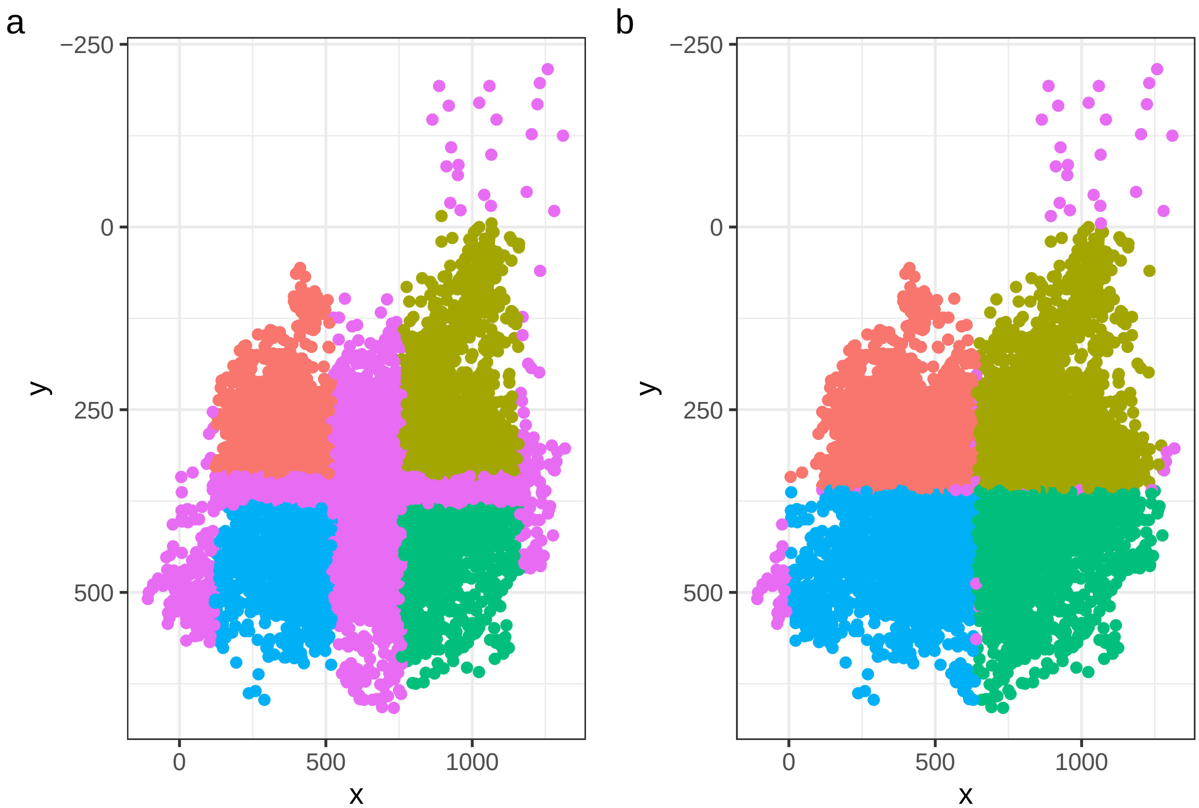


Figure 18. Example participant's gaze coordinates categorized into ROIs based on a) image bounding boxes and b) screen quadrants. Magenta points indicate looks that were not categorized into an ROI

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".

Methods

All stimuli, experiment scripts, data, and analysis scripts are available on the Open Science Framework at https://osf.io/eubsc/ (https://osf.io/eubsc/). The study

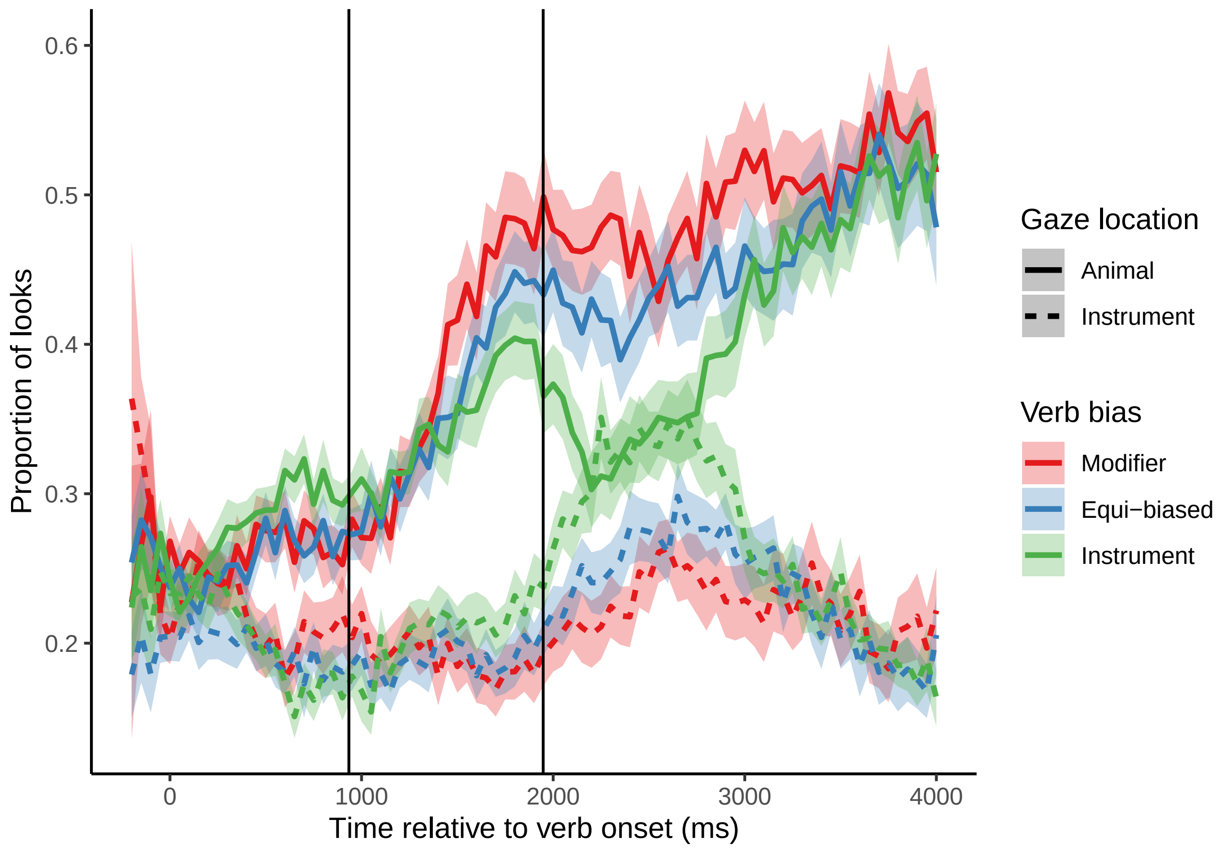


Figure 19. Timecourse of eye-gaze to target animal and target instrument by verb bias condition with gaze categorized based on which quadrant of the screen the coordinates fall in (as opposed to a bounding box around the image). Vertical lines indicate average onsets of animal and instrument offset by 200ms.

pre-registration is available at https://osf.io/tv57s (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 **??**).

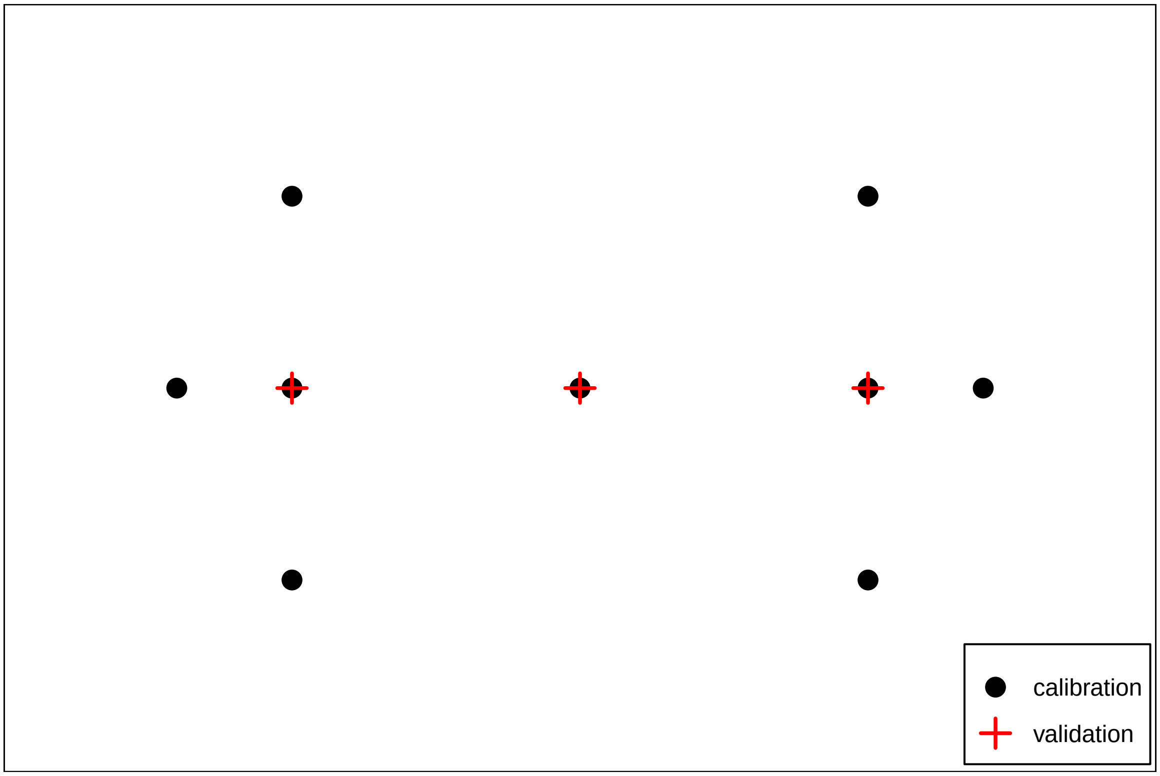


Figure 20. Calibration and validation point locations for Experiment 5. Black points were used for calibration. Red crosses were used for checking the accuracy of the calibration.

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 21). 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 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.

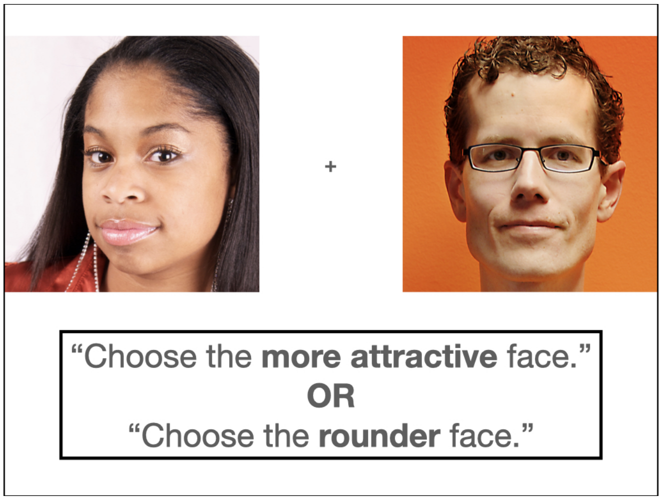


Figure 21. An example of a critical trial from Experiment 5. (Text did not appear on each screen.)

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 were paid $XX for completing the experiment. Data from 3 participants was excluded because their mode response made up more than 50% of their total responses , for a total of 24 participants in the norming. They each viewed all 172 faces and were asked to rate them on a scale from 1 (less attractive) to 7 (more attractive) using a slider. Faces were presented one at a time and in a random order for each participant. Following Shimojo et al., 19 face pairs were made by matching two faces that had a difference in mean attractiveness ratings that was 0.25 points or lower and that 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. **??**). 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.

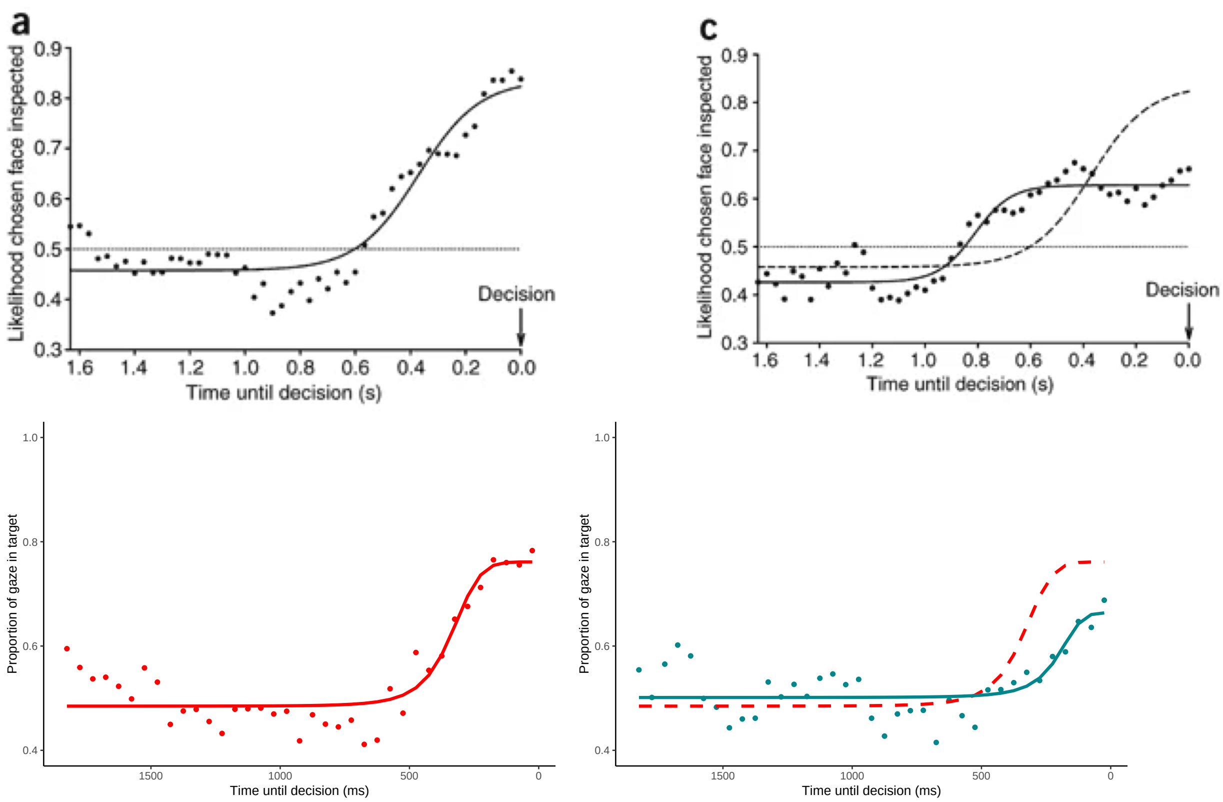


Figure 22. Primary results from Exp. 5. *Top* shows the original results from Shimojo and colleagues (Figures reprinted with permission[TODO]). The attractiveness judgment along with the best-fitting sigmoid is shown in the *top left*. Results for the roundness judgment are show in the *top right*, with 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.

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 is shown in Fig. 23.

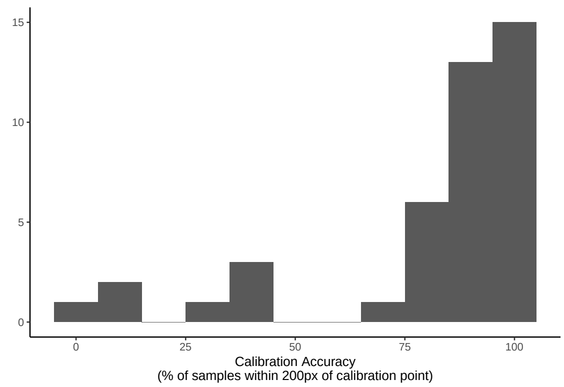


Figure 23. Histogram of calibration success in Exp. 5. Where participants required more than one calibration (N=8), only the final calibration was considered.

To determine whether calibration accuracy influenced our key effects, we calculated the percentage of samples during the task in which the participant was fixating the face they ultimately chose. There was a significant correlation for both the attractiveness judgments (r = 0.47 [0.04, 0.75], p = 0.03) and the roundness judgments (r = 0.60 [0.23, 0.82], p = 0). Inspection of Fig. 24 reveals that this correlation is due to a handful of particpiants with calibration values below 50%.

Thus, we re-analyzed the data, removing the participants whose calibration accuracy was not greater than 50%. This slightly improved the fits of the sigmoids (Attractiveness: $R^2 = 0.79$; Roundness: $R^2 = 0.60$). However, the difference between sigmoids remained non-significant using the Kolmogorov-Smirnov test (D = 0.22, p = 0.36). Descriptively, the results do not look substantially different (Fig. 25).

Effects of ROIs. In the original experiment, eye gazes that did not directly fixate one or other of the faces were excluded. In this section we explore an alternative coding of

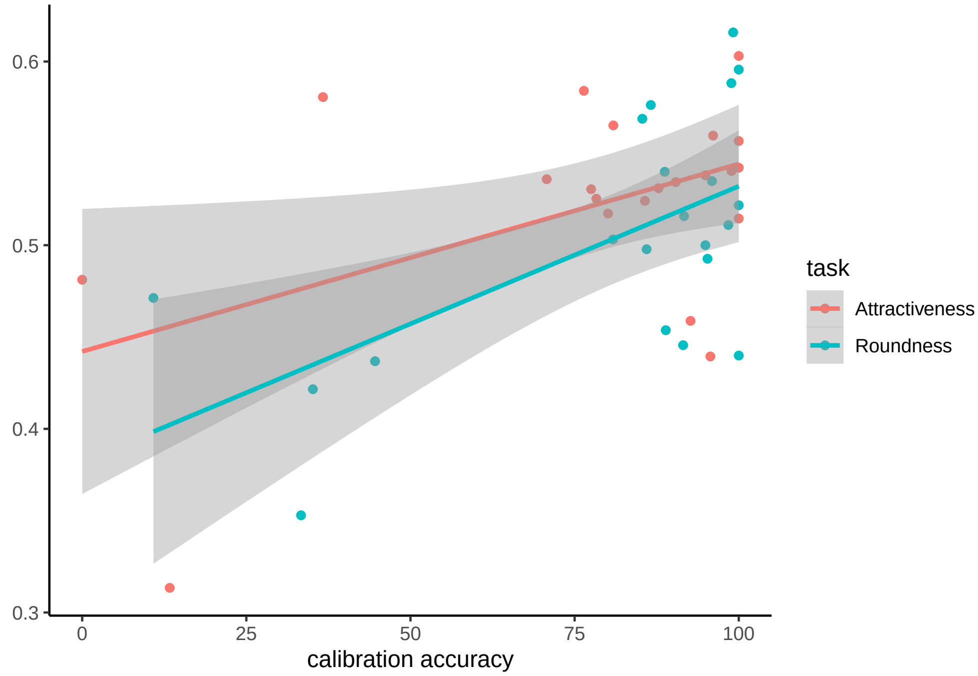


Figure 24. Correlation between calibration accuracy (x-axis) and percentage of samples fixating target (y-axis) in Exp. 5.

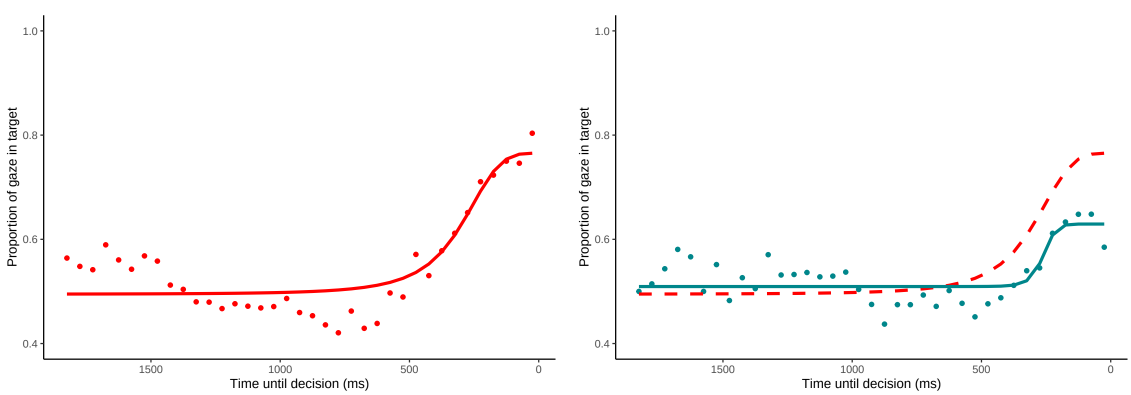


Figure 25. Revised results for Exp. 5 after removing low-calibration accuracy participants. *Left*: Eyegaze during attractiveness judgments, along with the best-fitting sigmoid. *Right*: Eyegze during roundness judgments, along with best-fitting sigmoid (best-fitting sigmoid for attractiveness is re-plotted with a dashed line for comparison).

the eye movement data by coding simply left half vs. right half of the screen. The coarser coding may be more appropriate for webcam-based eyetracking.

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

Discussion

Combined Analyses?

- Pooling data from all experiments we can look at patterns in the calibration and validation data

General Discussion

- E1:
- E2:
  - replication of key result with 4 quadrants
  - calib quality doesn't seem to matter
  - if attempting to control where people are looking, think about modifying task...
- E3:
- E4:
- E5:

References

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*(3), 238–299. https://doi.org/10.1016/j.cogpsych.2004.03.001

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

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.

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.