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

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

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

Author Note

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

Enter author note here.

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

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

Abstract

ADD LATER

Keywords: keywords

Word count: X

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

Intro stuff:

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

Present work

In the present work, we attempted to replicate six eye-tracking studies from the cognitive science literature using the eye-tracking plug-in from `jsPsych`, a Javascript library for running behavioral experiments in a web browser (de Leeuw, 2015). The plug-in relies on the `webgazer.js` library for webcam eye-tracking (Papoutsaki et al., 2016). The goal was to examine the strengths and weaknesses of webcam eye-tracking for common paradigms in cognitive science. The studies were chosen to cover a variety of topic areas (e.g., memory, decision-making, psycholinguistics) and paradigms (two halves of the screen, visual world paradigm with four quadrants, visual world paradigm with "natural" scenes). ...

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.

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 4.1.0; R Core Team (2021)] and the R-packages *afex* [Version 0.28.1; Singmann, Bolker, Westfall, Aust, and Ben-Shachar (2021)], *broom.mixed* [Version 0.2.6; Bolker and Robinson (2020)], *dplyr* [Version 1.0.7; Wickham, François, Henry, and Müller (2021)], *forcats* [Version 0.5.1; Wickham (2021a)], *ggplot2* [Version 3.3.4; Wickham (2016)], *jsonlite* [Version 1.7.2; Ooms (2014)], *lme4* [Version 1.1.27.1; Bates, Mächler, Bolker,

and Walker (2015)], *lmerTest* [Version 3.1.3; Kuznetsova, Brockhoff, and Christensen (2017)], *Matrix* [Version 1.3.3; Bates and Maechler (2021)], *papaja* [Version 0.1.0.9997; Aust and Barth (2020)], *readr* [Version 1.4.0; Wickham and Hester (2020)], *shiny* [Version 1.6.0; Chang et al. (2021)], *stringr* [Version 1.4.0; Wickham (2019)], and *tidyr* [Version 1.1.3; Wickham (2021b)] 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 XX 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 scrips, data, analysis scripts, and pre-registration are available on the Open Science Framework at https://osf.io/s82kz.

Participants. 60 participants were paid $XX 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 an -point eye-tracker calibration and validation. 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 had a sound understanding of 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.

Materials & Design. The visual stimuli were created through Canva and depicted an agent accompanied by four to five objects in the scene (see Figure XX). 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."

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.

TO DO: add figure

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). They examined how visuospatial information is integrated into memories 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

Participants. 60 participants were paid $XX for their participation . We analyzed the data of 59 participants, a number that was limited by budget constraints, but still 2.5x larger than the original sample size of 24, as suggested by Simonsohn (2015). We ended up

excluding the data of 1 participant whose eye tracking data seemed to be blank, rendering us unable to analyze it.

Procedure. The task began with an -point eye-tracker calibration and validation. 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 XX). The four possible categories were humanoids, household objects, animals, and methods of transportation. Participants were asked to remember the contents of the four quadrants. Different images were used in each block.

Each of the four quadrants was presented one at a time. First, a list of the items in the quadrant were shown, then the items in the actual quadrant were shown (??). For each item, an audio file would play ("???") asking the participant to use their arrow keys to identify which direction each item was facing (every item was facing either left or right (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. Then, after all four quadrants were presented in this way, the participant was 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 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.

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.

Data analysis.

Results

Replication.

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

Calibration.

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

Discussion

Experiment 3

The third study was a replication attempt of Manns, Stark, and Squire (2000) which aimed to show that the visual paired-comparison task, widely used in the patient literature,

tapped into declarative memory. In the visual paired-comparison task, two identical pictures were presented side by side for a brief viewing period. After a delay, one of the previously viewed pictures was presented along with a new picture. Individuals looked more at the new picture than the old picture and the time spent looking was correlated with later recognition memory performance. On the other hand perceptual priming, thought to recruit non-declarative memory, was not linked to later recognition. (The perceptual priming arm of the design was not included in this replication.)

Methods

Participants. Our initial sample size was 51 participants for the first day of our experiment and 48 of them came back for the second day. Following Manns et al., we excluded 3 participants due to perfect performance on the recognition memory test. 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. 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 pohotographs 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 XX

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.

Data analysis.

Results

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/ (https://osf.io/x3c49/). The pre-registration for the study can be found at https://osf.io/3v4pg (https://osf.io/3v4pg).

Participants. 58 (??) participants were paid $XX for their participation . A sample size of 58 was chosen because we wanted to replicate the experiment with greater statistical power. Note that the original study had a sample size of 24.

Procedure.

- *TO DO:* add details of calibration point locations

After the eye-tracking calibration, 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 XX)[1]. Using their cursor, participants could act out the instructions by clicking on objects and moving them or motioning over the objects[2]. After the action was completed, the participants were instructed to press the space bar which led to a screen that said "Click Here" in the middle in order to remove bias in the eye and mouse movements from the previous trial. The experiment only allowed the participants to move on to the next trial once the audio was completely done playing and the mouse had been moved over at least one object.

TO DO: ADD FIGURES Figure 1: An example of a critical trial 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.

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

[1] 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

[2] 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.

instrument combinations (e.g., "rub" could be paired with "panda" and "crayon" in one list and "panda" and "violin" in the second list). Within each list, the same verb was presented twice but each time with a different target instrument and animal. The lists were randomly assigned to the participants to make sure the effects were not caused by the properties of the animal or instrument images used. 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 1 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 entered as varying intercepts with by-participant varying slopes for verb bias condition[3]. 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, $p < 0.01$)

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

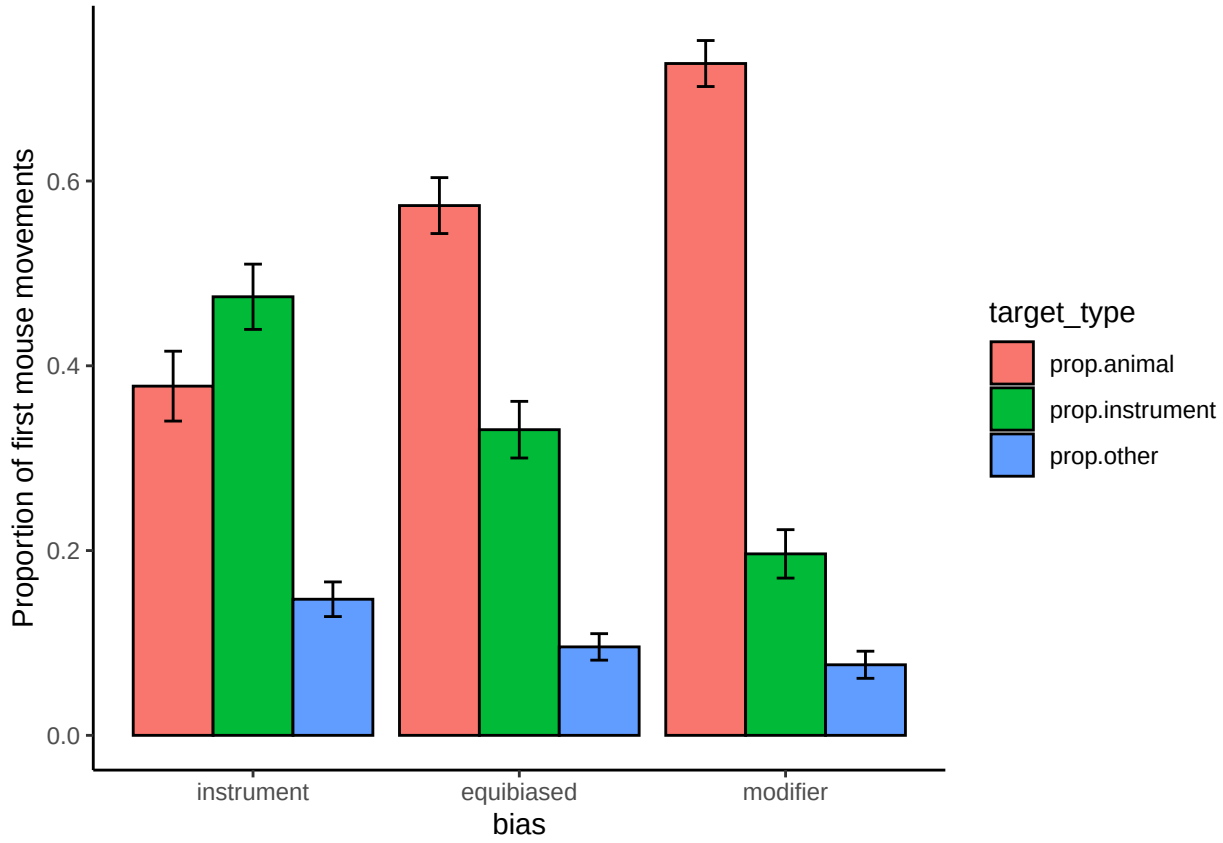

Figure 1. Proportion of first mouse movements by location and verb bias.

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 2 suggests that the participants made more fixations to the target animal when the verb was modifier-biased. . .

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

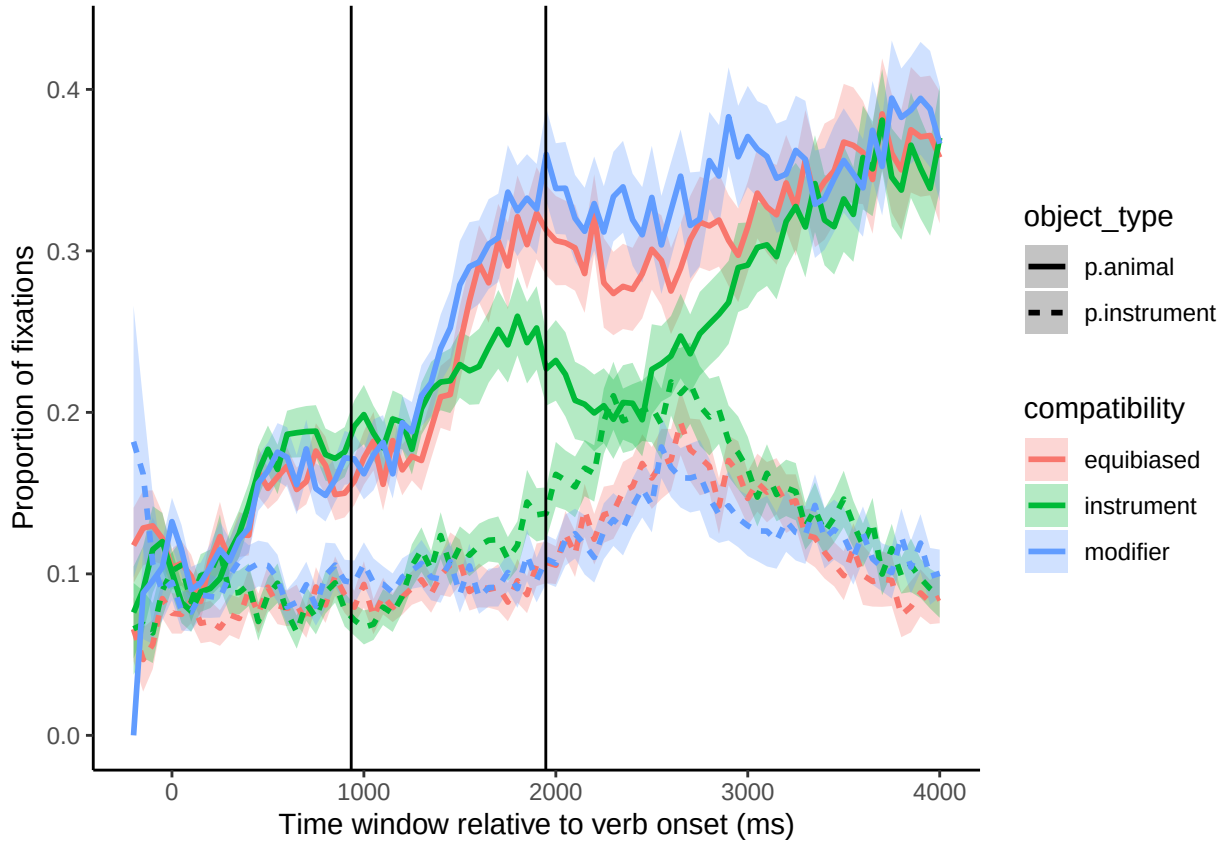

Figure 2. Timecourse of eye-gaze to target animal and target instrument by verb bias condition. vertical lines indicate average onsets of animal and instrument offset by 200ms.

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[4]. 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 relative to the equi-biased and instrument-biased condition (b = 0.03, SE = 0.01, $p < 0.01$) and in the equi-biased relative

[4] `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.

to the instrument-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 relative to the equi-biased and instrument-biased condition ($b = 0.08$, $SE = 0.02$, $p < 0.01$) but not significantly so in the equi-biased relative to the instrument-biased condition ($b = 0.03$, $SE = 0.02$, $p = 0.59$).

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 level of the individual

Effects of ROIs.

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

Discussion

Experiment 5

The fifth study was a replication attempt of Shimojo, Simion, Shimojo, and Scheier (2003), which found that human gaze is actively involved in preference formation. Participants were shown pairs of human faces and were asked to choose the face which they found more attractive. Prior to making their explicit selection, participants spent more time looking at the face that was ultimately judged more attractive. This bias was not as large when the task was face shape discrimination rather than a preference-based choice.

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

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

Participants. 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.

50 participants for the main task were recruited on Prolific and were paid $XX. 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.

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 XX). 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.

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. Participants in the norming study viewed 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, and eye position was then tracked using MediaAnalyzer software. This eye tracker was able to collect 30 samples per second. In our study, the camera on the device used by the participant and the jsPsych WebGazer package were used to determine where a participant's gaze was directed. Since our eye tracking method did not have the same sample collection rate as the eye tracker in Shimojo et al., time windows of 50 milliseconds were used when looking at the likelihood of gazing at the chosen face at a given point between a cutoff time and when the decision was made.

Results

Discussion

Experiment 6

The sixth study was a replication attempt of Posner et al. ??.

Methods

Participants.

Procedure and Design.

Data analysis.

Results

Replication.

Calibration.

ROIs.

Item Numbers.

Discussion

Combined Analyses

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

General Discussion

References

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

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

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

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

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. In *Proceedings of the 25th international joint conference on artificial intelligence (IJCAI)* (pp. 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

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