Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/7/eaav2558/DC1

Supplementary Materials for

The origin of pointing: Evidence for the touch hypothesis

Cathal O'Madagain*, Gregor Kachel, Brent Strickland

*Corresponding author. Email: cathalcom@gmail.com

Published 10 July 2019, *Sci. Adv.* **5**, eaav2558 (2019)
DOI: 10.1126/sciadv.aav2558

This PDF file includes:

Fig. S1.

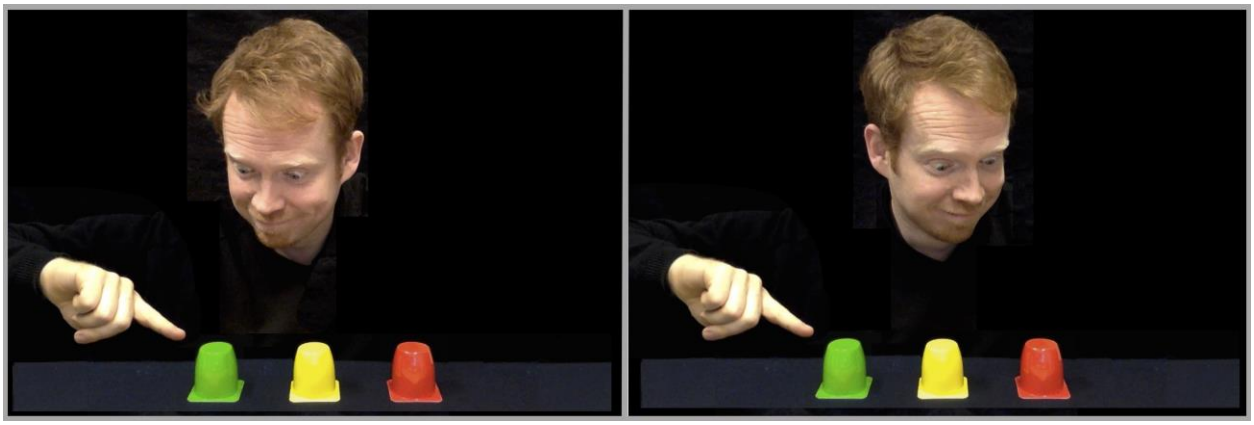

Fig. S1. Interpretation experiment, controls. On the left is control 1. Here the finger-tip is closest to the cup the figure is gazing at (green), but the 'arrow' also picks out that cup. We expected all participants to reliably pick the green cup in their interpretation of this 'no conflict' condition, since this should be the result whether the 'arrow' or 'touch' interpretation is applied. Participants who did not pick the cup the figure is gazing at in this condition in 3 of 4 trials were excluded from the study (no participants were excluded for this reason). On the right is control 2, where the gaze conflicts with both arrow and the cup the finger-tip is closest to touching; our concern here was to determine whether participants were simply following the figure's gaze, in which case they should have ignored the location of the fingers (something that turned out only to be a possible interpretation of the 6yr-olds). However, no group followed gaze above chance in this condition, ruling out this interpretation (see fig. S2) (Photo credit: C. O'Madagain, Max Planck Institute for Evolutionary Anthropology).

Fig. S2.

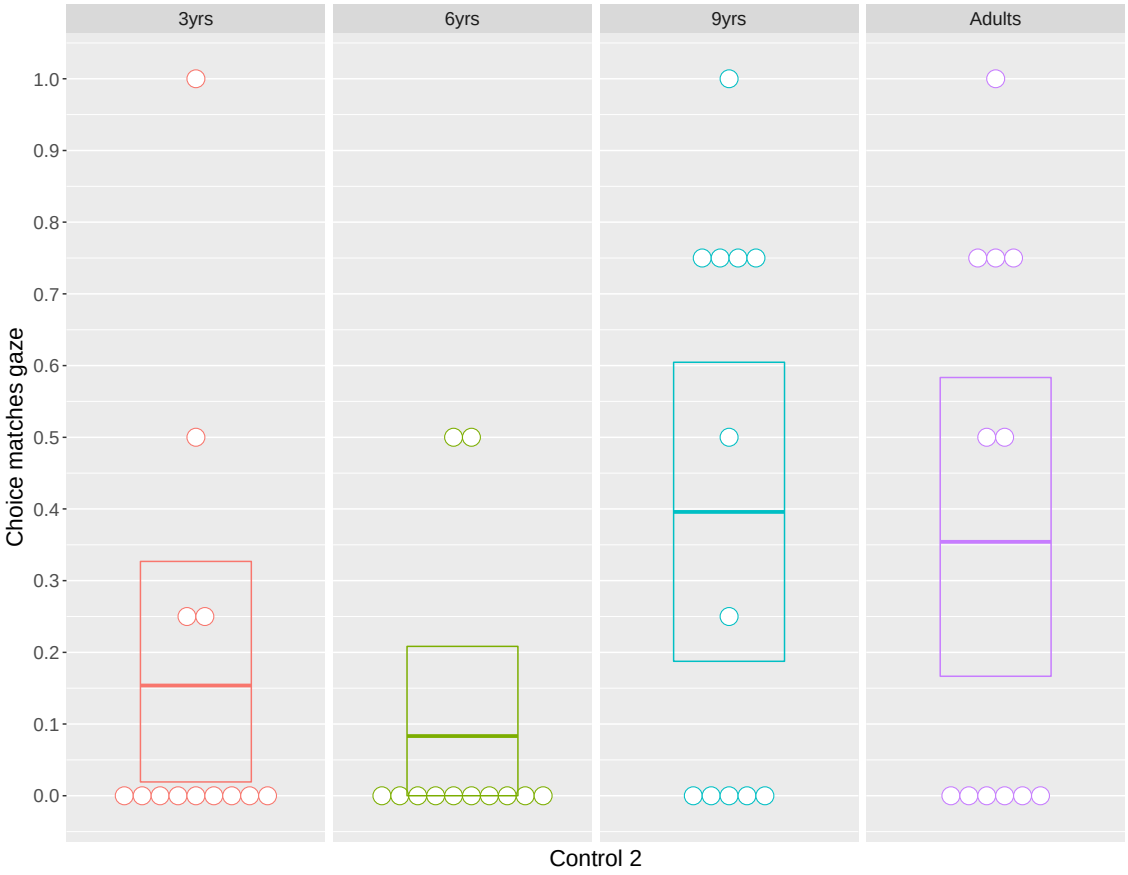

Fig. S2. Interpretation experiment—results of control 2. Since no age group was above chance for following the figure's gaze in this control condition, this shows that the results cannot be explained for any group merely by gaze following alone. We did not run this control for the 18mth-olds, but since they do not follow gaze in the 'arrow' condition, gaze following alone cannot explain their results.

Details on Materials and Methods

General note on Participants in all Studies:

Participants were recruited from the Leipzig area. In the case of 18mths, 3yr, 6yr olds, 9yr olds, parents were invited to come with their child to the Max Planck Institute for Evolutionary Anthropology in Leipzig. Children were rewarded for their participation with a soft toy and a t-shirt. Adult participants were recruited from the staff of the Max Planck Institute. The same participants took part in all three studies in age groups 3yrs, 6yrs, 9yrs and adults, except where participants were dropped (as noted below on a case by case basis), in which case new participants were used as replacements who did not participate in all experiments. The order of participation was counterbalanced, so that each participant was assigned to one of six possible orders of the three experiments. Different 18mth-olds were used for each experiment, given attention limits.

Study I: Reference-Fixing Experiment

Piloting:

We piloted six 3yr-olds with an early version of the procedure, and found that the 'touch-line' was clearly a better predictor of reference than the arrow-line. As a result, we specified the number of participants in advance at the smallest number we believed we could find reliable statistical results for while properly counterbalancing the stimuli, so that we could run multiple age-groups. We settled on twelve participants per age group for this and all subsequent experiments. We ran further pilots to refine the procedure (eleven 18mth olds, eleven 3yr olds, five 6yr olds, and five adults).

Final participants:

Final participant numbers varied slightly by age group (e.g. the 3 year olds) due to overscheduling and our laboratory policy not to turn away participants; the variation is taken into account in the model. Thirteen adults (mean=29.06 years, 9.55years, 18.24-42.42years; 7 female); twelve 6-year-olds (mean=73.24mths; SD=19days; range=71.50-73.70); seventeen 3-year-olds (mean=36.21; SD=12 days; range=35.34-36.78; 9 female), and thirteen 18-month-olds (mean age=18.62mths; SD=0.27mths; range=18.09-18.93mths) were included in the analysis.

Dropped Participants:

Two adults were excluded because they revealed after the test that they were familiar with the hypothesis (they had heard the first author present the idea at a talk); two 3-year-olds were excluded, one refused to participate and the other due to a camera failing to record the session; five 18-month-olds were excluded for participating in fewer than four trials.

Dropped Trials:

The main reason for dropping trials was that the participant's hand or eye were out of view in the recording, or in the version run for 18mth-olds, if the target was off-camera when it appeared from behind the curtain. Since this task required us to draw a vectors including the eye, finger-tip, and target trials in which one of these were obscured were impossible to code. We also dropped trials where the participant was not in the position we had asked them to sit in, for example by standing

up or crawling onto the table toward the targets. If the gesture was produced very quickly it sometimes resulted in a blur in the screenshot, which again could not be measured. We also dropped trials if a parent interfered with the child's gesture, or if a participant was confused about the task and it was not clear to the experimenter that s/he was pointing at one of the target cups.

Below are numbers of dropped trials by age group, with the reasons listed afterwards, and number of trials for each reason in parentheses.

Adults: 10 of 284 trials, or 3.85%. Reasons: participant out of position (1); gesture produced too quickly to code (2); participant looking away from target (5); eye out of view (2).

6-year-olds: 21 of 323 or 6.5%. Reasons: participant is confused about task (14); participant looking away from targets (3); eye out of view (2); gesture produced too quickly to code resulting in a blur (2).

3-year-olds: 99 of 438 trials, or 22.8%. Reasons: eye out of view (20); participant out of position (4); participant looking away from targets (36); gesture produced too quickly resulting in a blur (8); unclear to experimenter if child is pointing at targets (31); parent interferes (1).

18-month-olds: 18 of 161 trials, or 11.18%. Reasons: gesture produced too quickly resulting in a blur (11); target off camera (2); eye out of view (5).

Reliability Coding:

A second coder blind to the hypothesis followed the coding for one third of the participants in each age group (following standard protocol in our laboratory). The correlation for the judgments of the time in the video when the gesture is most fully produced was high and significant (rho 0.99, $p < 0.001$), as was the correlation for the angles between the vectors and the target was high and significant (rho 0.95, $p < 0.001$).

Analysis of Reference-Fixing Experiment:

In order to test whether one vector was a better predictor of the reference of a pointing gesture and whether this varied between ages we used a Generalized Linear Mixed Model with binomial error structure and logit link function. The response in this model was whether, in a given trial, the touch-line was a better predictors of a participant's pointing style or not (when both touch-line and finger-line were equally good we excluded the trial from the data). As the sole test predictor (*25*) we included age of the participants (factor with four levels) and we also controlled for trial number (both of which were included as fixed effects). As a random effect we included the ID of the participants, and to keep type I error rate at the nominal level of 0.05 we included random slopes (*24*) of trial number within participant but not the correlation between the random intercept and slope.

To test whether the probability to point along the touch-line was larger than chance (0.5) we tested the intercept of the model. In order to do so we manually dummy coded age group and then centered the derived dummy variables to a mean of zero and also z-transformed trial number to a mean of zero and a standard deviation of one. We then added a constant to the predictors such that the logit transform of the intercept was exactly equal to the average probability of pointing

according to the touch-line (needed because in a logistic model centering of the predictors alone does not ensure the intercept to exactly model the average response) and finally tested the Intercept using Wald's z-approximation (*21*). the sample size for this model was a total of 991 observations of 57 subjects.

To test the extent to which the two angles (finger angle and eye angle) differed and whether this was influenced by age we used a linear mixed model (LMM) with Gaussian error structure and identity link. The response was the difference between the finger and the eye angle in a given trial (in degrees) and age was the only test predictor with fixed effect; but to control for potential learning effects we also included trial number (z-transformed). The random effects structure was the same as in model 1, and sample size for this model was 1057 angle differences for 57 subjects.

Study II: Rotation Experiment

Piloting:

The procedure was developed with pilot subjects, including eighteen 3yr-olds, four 6yr-olds, and four 18mth-olds.

Final participants:

Thirteen adults (mean age=29.06yrs, SD=9.55yrs; range=18.24-42.42yrs; 7 female); twelve 6-year-olds (Mean age=73.24mths; SD=19days; range=71.50-73.70mths); and sixteen 3-year-olds (mean age=36.21mths; SD=12 days; range=35.34-36.78mths; 9 female) were included in the analysis. In the version of the study we ran for 18-month-olds a further 13 children were included (mean age=18.52mths; SD=9 days; range=18.03-18.88mths).

Dropped Participants:

Two adults were excluded and replaced because they revealed that they were familiar with the experimental hypothesis, as in the first experiment; four 3-year-olds refused to participate; and eight 18-month-olds were excluded for participating in fewer than four trials.

Dropped Trials:

Adults: 8 of 316 trials or 2.5%. Reasons: participant confused about task (4); gesture produced too quickly to code, creating a blur (2); camera out of focus (2).

6yrs: 1 of 393 trials or 0.2%. Reasons: participant out of position (1).

3yrs: 4 of 482 or 0.8%. Reasons: participants pointed with both hands together at different rotations, making it impossible to give a single rotation measure (4).

18mths: 15 of 120 or 12.5%. Reasons: participant confused about task (14); gesture produced too quickly to code, resulting in blur (1).

Reliability Coding:

A second coder followed the coding procedure for a third of the participants in the sample. Correlation on choice of frame was high and significant (rho 0.99, $p < 0.001$), and agreement on hand rotation (0-4) was good (Kappa = .74).

Analysis of Rotation Experiment:

To test whether the rotation of the wrist corresponded to the side at which the target was presented and whether the degree of correspondence varied with age and or condition (2D or 3D) we fitted a GLMM with binomial error structure and logit link function. For this test we considered only trials in which the target was presented to the left or right. The response coded whether the rotation direction (left of right) matched the side where the target was presented. As fixed effects we included condition (2D or 3D) and its interaction with age. We further controlled for age, object position and the hand used (and their interaction) and trial number. As a random effect (intercept) we included the identity of the tested individual, and we further included random slopes of condition (2D or 3D), trial number, object position, the hand used, and the interaction between the latter two (all categorical predictors entered as random slopes were manually dummy coded and then centered to a mean of zero). The null model (*25*) lacked condition and its interaction with age but was otherwise identical to the full model. The sample size for this model was 703 pointing events by 55 subjects.

Study III: Interpretation Experiment

Piloting:

This was a difficult procedure to refine, and we developed it with a relatively large group of pilot subjects, including ten 12mth-olds (who we found unable to participate), fifteen 18mth-olds, forty-three 3yr-olds, eleven 6yr-olds, two 9yr-olds, and eleven adults.

Final participants:

Twelve adults (mean age=29.47yrs, SD=9.86yrs, range=18.24–42.42yrs; 6 female); twelve 9-year-olds (mean age=108.16mths; SD=39days; range=106.13-109.93mths; 6 female); twelve six-year-olds (mean age=73.24mths; SD=19.61days; range=71.50-73.70mths; 6 female); and twelve 3-year-olds (mean age=36.25mths; SD=12 days; range=35.34–36.78mths; 6 female); 18-month-olds included 24 participants in a between participant design (mean age=18.4mths, SD=7days, range=18.03–18.81mths; 11 female). We chose to include 9-year-olds in this task because the 6-year-olds were at chance in their interpretation between the 'arrow' and 'touch' hypotheses, and we hypothesized that older children might lean toward the arrow interpretation, as indeed they did.

Coding and Reliability:

The main measure is which cup (in the version of the study run with 3-year-olds, 6-year-olds and adults), or box (for 18-month-olds) was chosen, either by naming the color of the cup, or removing the cloth cover of the box. The first cup or box that a participant chose counted as their decision. A second coder replicated the coding for 42% of the sample, and reliability was very good (18mths: Kappa=.87; 3yrs, 6yrs, adults: Kappa=.98).

Dropped Participants:

Three adults were excluded: two told us afterward they were familiar with the hypothesis (they had heard the first author present the idea at a talk), and one told us after the test that she had been distracted and was not looking at the slides while answering the questions; eight 3-year-olds were excluded: four did not know the color terms, three completed less than four trials; and in one case the camera malfunctioned; eight 18-month-olds were excluded, for completing less than four trials each.

Dropped Trials:

18mths: 15 of 281 or 5.3%. Reasons: Participant chose both objects by dragging on both cloth covers (9); participant was not attending to experimenter before choosing (2); participant made no choice (4).

Reliability Coding:

A second coder replicated the coding for 42% of the sample, and reliability was very good (18mths: Kappa = .87; 3yrs, 6yrs, adults: Kappa = .98). The first cup or box that a participant chose counted as their decision.

Analysis of Interpretation Experiment:

To model participants reliability in choosing the cup matching gaze in each condition, we used a Generalized Linear Mixed Model with binomial error structure and logit link. The key predictors with fixed effects were condition, subject age and their interaction. We also included trial number as a further fixed effect, as a control. Participant ID was included as a random effect (intercept) and we included random slopes of trial number and gaze direction (manually dummy coded and then centered to a mean of zero). The null model lacked gaze direction, subject age and their interaction but was otherwise identical to the full model. Sample-size for this model was 1010 trials of 72 subjects.

Analysis: General Considerations and Implementation

We fitted the models in R (version 3.3.1; (*19*)) using the functions lmer (LMMs) or glmer (GLMMs) of the R package lme4 (version 1.1-12; (*20*)). In case of LMMs we ensured normality and homogeneity of residuals by visual inspection of QQ-plots (*21*) and residuals plotted against fitted values (*22*) which revealed no serious violations of these assumptions. We established model stability by excluding subjects one at a time and comparing model estimates derived from models of the respective subsets with those obtained for the full data set which revealed no influential subjects to exist. We tested for the effect of individual fixed effects by comparing a full with a respective reduced model lacking the fixed effect (but being otherwise identical) using a likelihood ratio test (*23*; *24*). We derived confidence intervals by means of a parametric bootstrap (function bootMer of the package lme4). In the case of the data in the Rotation experiment, but not the others, we made the bootstrap conditional on the particular individuals tested (argument use.u set to TRUE). The reasoning behind this conditional bootstrap was that the probabilities of wrist rotations matching

the side of the box varied greatly between subjects, leading to unrealistic results for the random effects and consequently very wide confidence intervals.