

1 Supplements for the manuscript ‘TANGO: A reliable, open-source, browser-based task to
2 assess individual differences in gaze understanding in 3 to 5-year-old children and adults’

3 Julia Prein¹, Manuel Bohn¹, Steven Kalinke¹, & Daniel B. M. Haun¹

4 ¹ Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary
5 Anthropology, Leipzig, Germany

6 Author Note

7 Correspondence concerning this article should be addressed to Julia Prein, Max
8 Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig,
9 Germany. E-mail: julia_prein@eva.mpg.de

10 Supplements for the manuscript ‘TANGO: A reliable, open-source, browser-based task to
 11 assess individual differences in gaze understanding in 3 to 5-year-old children and adults’

12 **Effects of trial type and trial number**

13 Children showed nearly perfect precision in the first training trial. As visual access to
 14 the target location decreased in the succeeding training trials, imprecision levels increased.
 15 Within test trials, children’s imprecision levels did not vary as a function of trial number.

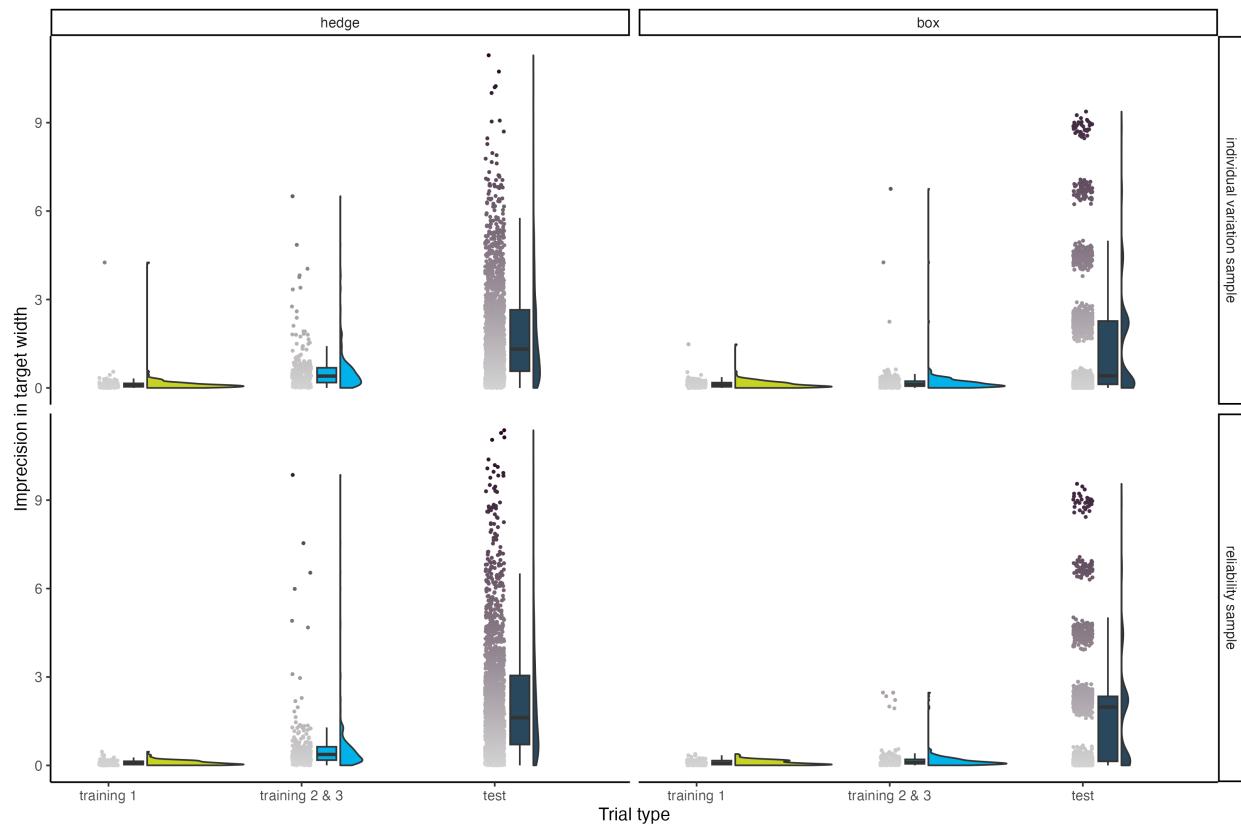


Figure 1. Imprecision by trial type, split by study version and sample. The x axis represents the trial type. The y axis represents imprecision, i.e., the absolute distance between the target’s center and the participant’s click. The unit of imprecision is counted in the width of the target, i.e., a participant with imprecision of 1 clicked one target width to the left or right of the true target center. Small dots show the imprecision for each subject in each trial. Boxplots (boxes represent first to third quartiles of data; vertical lines indicate the median; horizontal black lines display the range) and a half violin plot show the data distribution.

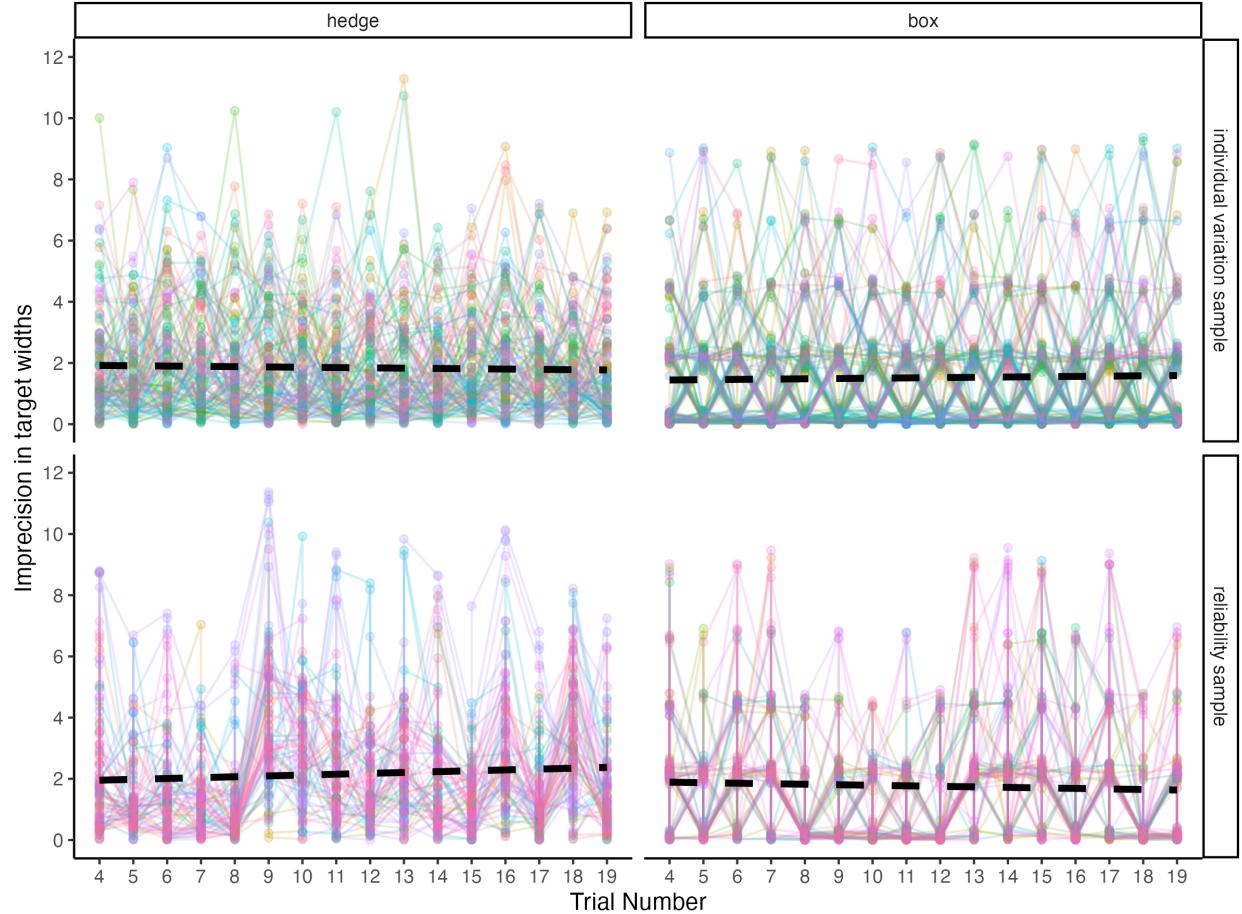


Figure 2. Imprecision across test trials, split by study version and sample. The x axis represents trial number. The y axis represents imprecision, i.e., the absolute distance between the target's center and the participant's click. The unit of imprecision is counted in the width of the target, i.e., a participant with imprecision of 1 clicked one target width to the left or right of the true target center. The black dashed regression lines show smooth conditional means based on linear models. Small colored dots show the imprecision for each subject in each trial. Colored lines connect the trials of each individual.

16

Webcam coding of the child sample

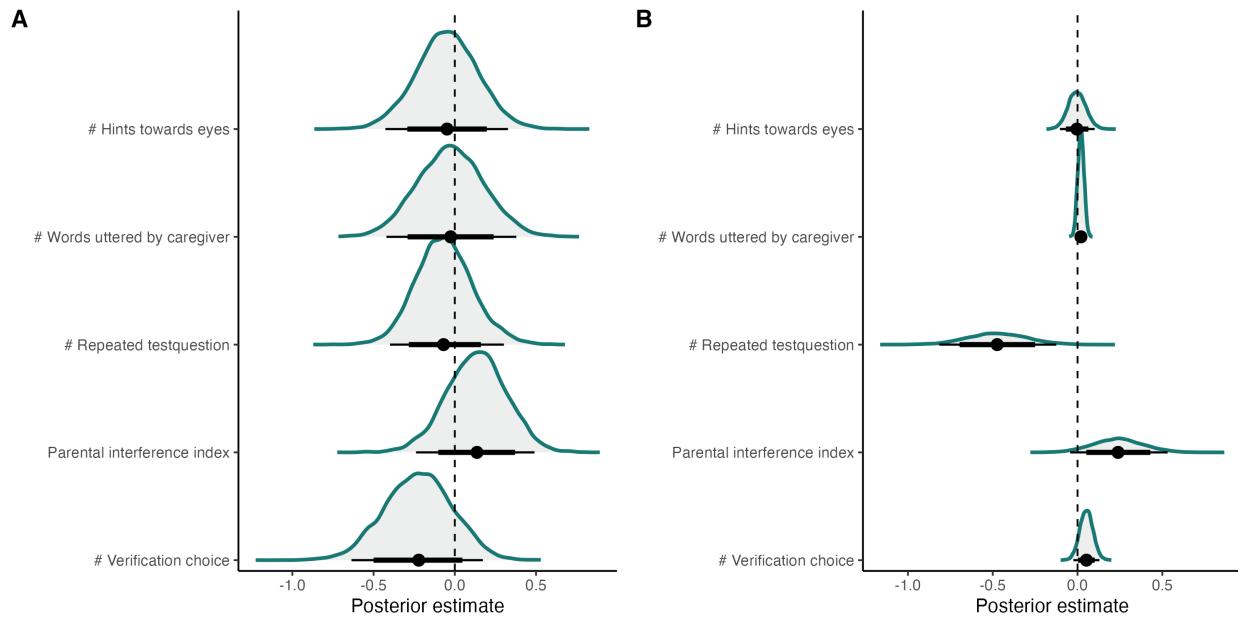


Figure 3. Model comparison for exploratory webcam coding of parental interference. Factors of parental interference and their influence on the probability of responding correctly. The graph shows the estimated density curves of a model's predictor coefficient. Models are ordered according to their WAIC scores in the trial-by-trial analysis, with the uppermost winning the model comparison. (A) Analysis on a trial-by-trial level. (B) Analysis on a subject level (i.e., average across trials per subject).

17 Comparing the performances of children across our two data collection modes, we
 18 found that children participating remotely were slightly more precise. This difference was
 19 especially prominent in younger participants in the box version of the task. It is
 20 conceivable that caregivers were especially prone to influence the behavior of younger
 21 children. In the box version, caregivers might have had more opportunities to interfere
 22 since they carried out the clicking for their children. In an exploratory analysis, we coded
 23 parental behavior and environmental factors during remote unsupervised testing. Due to
 24 the time consuming nature of hand coding videos frame by frame, we focused on the
 25 subsample with the greatest performance difference between data collection modes: the
 26 three-year-olds in the box version of the task ($n = 16$). We reasoned that if parental
 27 interference cannot explain the greatest performance difference in our sample, the effects

would be negligible in the remaining sample. A trial was defined as the time between two eye blinking sounds. We transcribed all utterances by parents and children and counted the words uttered by each. We then classified the utterances into several categories: question asked by child, repeated test questions by caregiver, hints towards agents (how many times the caregivers guided the child's attention to the agent), hints towards eyes (how many times the caregivers guided the child's attention to the agent's eyes), verification of choice (how many times the caregiver questioned or double checked the child's response), mentioning of screen (how many times the caregiver verbally guided the child's attention to the screen), pointing to screen (how many times the caregiver pointed towards the screen), positive & negative feedback, motivational statements, and incomprehensible utterances.

In addition, we coded how many adults and children were present, whether a response click was obviously conducted by the caregiver themselves, and whether children took a break during the trial. We conducted a model comparison to estimate the effects of parental interference. Our null model explained the response behavior by age, while including random effects for subject and target position (model notation in R: `correct ~ age + symmetricPosition + (1 + symmetricPosition | subjID)`).

We compared this null model to models including the number of words uttered by the caregiver, number of repeated testquestions, verification of choice, or hints towards eyes as fixed effects. Furthermore, we calculated an parental interference index by summing up number of repeated testquestions, verification of choice, and hints towards eyes, with the sign matching the variable's direction of effect. Remaining variables that we coded for were not included since there was not enough variation and/or occurrences in our sample. We compared models using WAIC (widely applicable information criterion) scores and weights. As an indicator of out-of-sample predictive accuracy, lower WAIC scores stand for a better model fit. WAIC weights represent the probability that the model in question provides the best out-of-sample prediction compared to the other models. On the trial level, the model including the verification of choice as a main effect performed best: here, the less the

55 caregivers asked for children's responses again, the more likely children clicked on the
 56 correct box. Interestingly, the effect reversed on a subject level - possibly due to greater
 57 learning effects for the children that were most likely to click incorrectly in the beginning
 58 and then receiving most parental comments. On the subject level, the model including
 59 number of repeated test questions performed best: the more caregivers asked again where
 60 the target landed, the more likely children were to respond to the incorrect box. In all
 61 cases, however, ELPD difference scores were smaller than their standard errors. Similarly,
 62 95% CI of the model estimates included zero and were rather wide. Therefore, we conclude
 63 that the effect of parental interference was negligible and could, most likely, be explained
 64 as described above.

Predictor	WAIC	SE_WAIC	Weight	ELPD_DIFF	SE_ELPD_DIFF
By trial - # Verification choice	255.98	15.66	0.34	0.00	0.00
By trial - Null model	256.77	15.05	0.23	-0.40	1.18
By trial - Parental interference index	257.24	15.32	0.18	-0.63	0.99
By trial - # Repeated testquestion	258.57	15.25	0.09	-1.30	1.23
By trial - # Words uttered by caregiver	258.88	15.37	0.08	-1.45	1.13
By trial - # Hints eyes	258.88	15.35	0.08	-1.45	1.21
By subject - # Repeated testquestion	83.61	10.23	0.81	0.00	0.00
By subject - Parental interference index	88.27	12.17	0.08	-2.33	3.58
By subject - Null model	89.16	11.21	0.05	-2.78	3.60
By subject - # Verification choice	89.96	12.04	0.03	-3.18	4.47
By subject - # Words uttered by caregiver	90.56	10.70	0.02	-3.48	4.23
By subject - # Hints eyes	93.02	11.84	0.01	-4.71	3.76

65

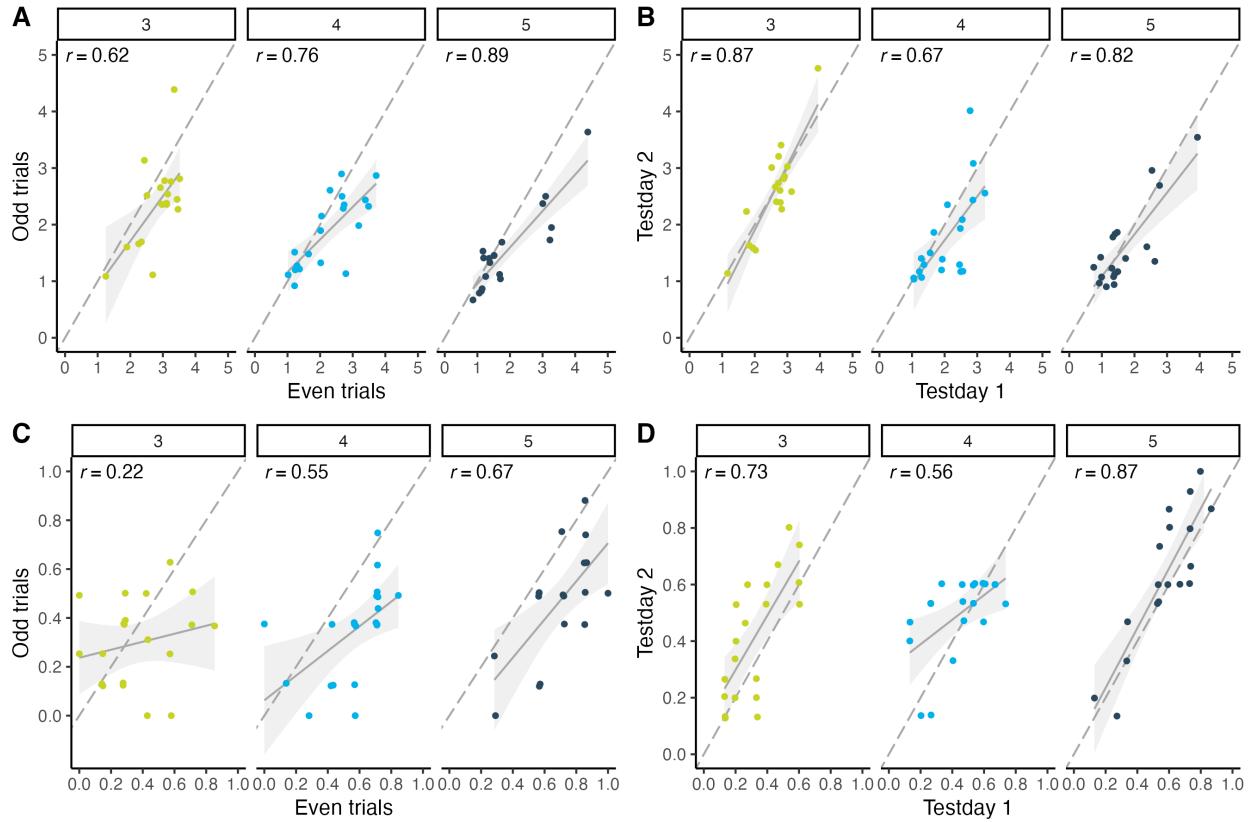
Reliability by age group

Figure 4. Reliability split by age group. (A) Internal consistency (odd-even split) in hedge child sample by age group. (B) Test-retest reliability in hedge child sample by age group. (C) Internal consistency (odd-even split) in box child sample by age group. (D) Test-retest reliability in box child sample by age group. For the hedge version, performance is measured as imprecision, i.e., the absolute distance between the target's center and the participant's click (averaged across trials). The unit of imprecision is counted in the width of the target, i.e., a participant with imprecision of 1 clicked on average one target width to the left or right of the true target center. For the box version, performance is measured as the proportion of correct responses, i.e., how many times the participant clicked on the box that contained the target. Regression lines with 95% CI show smooth conditional mean based on a linear model (generalized linear model for box version), with Pearson's correlation coefficient r . Dots show the performance for each subject. The color of data points denotes age group.

66

Validity

67 Social-environmental factors

68 **Participants.** We included all children where families filled out a short
69 demographic questionnaire. This subsample consisted of 109 children, including 30
70 3-year-olds (mean = 41.80 months, SD = 3.44, range = 36 - 47), 36 4-year-olds (mean =
71 54.19 months, SD = 3.08, range = 48 - 59), and 43 5-year-olds (mean = 66.28 months, SD
72 = 3.62, range = 60 - 71).

73 **Analysis.** To estimate social environmental influences on gaze understanding, we
74 fitted GLMMs predicting the task performance by each of our questionnaire variables,
75 controlling for age (scaled), data collection mode (reference category: remote unsupervised)
76 and study version (reference category: box version): `cor_tango | trials(n_tango) ~`
77 `age + datacollection + studyversion`). In order to combine data of our two study
78 versions, we transformed continuous click responses from the hedge version into a discrete
79 outcome. For the target position, we categorized two adjacent bins as one imaginary box.
80 To measure participants' performance, we created imaginary box boundaries around the
81 target's landing position and examined whether the participant's click response fell into
82 this imaginary box. Across the two study versions, we could consequently model the
83 number of participant's correct responses using a *Binomial* distribution. For model
84 comparisons, we ran separate models, each with one of the following predictors as a fixed
85 effects added to the null model: number of household members, number of children aged
86 0-18 in household, number of children aged 1-12 in household, hours spent in childcare each
87 day, and age when subject entered childcare. In addition, we calculated three index scores.
88 First, we calculated a sibling variety score according to Peterson (2000). Second, we
89 implemented the modified version of Cassidy, Fineberg, Brown, and Perkins (2005). Third,
90 based on our own data exploration, we calculated the amount of peer exposure determined
91 as the number of siblings and the average hours spent in childcare. We compared the

models using WAIC (widely applicable information criterion) scores and weights (McElreath, 2020). As an indicator of out-of-sample predictive accuracy, lower WAIC scores stand for a better model fit. WAIC weights represent the probability that the model in question provides the best out-of-sample prediction compared to the other models.

Predictor	WAIC	SE_WAIC	Weight	ELPD_DIFF	SE_ELPD_DIFF
Age of childcare entry	571.66	28.13	0.73	0.00	0.00
Null model	575.85	30.39	0.09	-2.09	4.49
Sibling variety score (Peterson, 2000)	577.91	29.85	0.03	-3.12	4.64
# Children in household aged 0-18	578.01	29.70	0.03	-3.18	4.72
Peer exposure index	578.04	30.23	0.03	-3.19	4.70
Sibling variety score (Cassidy et al., 2005)	578.36	29.88	0.03	-3.35	4.65
# Household members	578.70	30.35	0.02	-3.52	4.50
# Children in household aged 1-12	578.93	30.04	0.02	-3.64	4.86
Average hours spent in childcare per day	579.03	30.69	0.02	-3.69	4.57

Results. Note that we did not find a great difference in WAIC scores between the compared models (see Supplements for WAIC scores and weights). The model estimates were all considerably smaller than estimates of age, study version and data collection, and all 95% CIs included zero. Nevertheless, a general pattern emerges: exposure to a more variable social environment positively influenced children's gaze understanding. The number of people and, more specifically, children, as well as the more diverse their age, the more likely children were to understand the agent's gaze cue. The only predictor resulting in a negative estimate was the age at which a participant entered childcare, i.e., the later a child entered, the better performance in the task.

Receptive vocabulary

Participants. Our sample consisted of 117 children, including 11 3-year-olds (mean = 45.05 months, SD = 1.58, range = 43 - 47, 26 4-year-olds (mean = 53.17 months, SD = 3.01, range = 48 - 57), 32 5-year-olds (mean = 65.80 months, SD = 3.23, range = 61 - 71), 36 6-year-olds (mean = 78.36 months, SD = 3.26, range = 72 - 84), and 12 7-year-olds (mean = 88.11 months, SD = 1.96, range = 84 - 91).

111 **Materials.** We employed the oREV, an Item Response Theory based open

112 receptive vocabulary task for 3 to 8-year-old children (Bohn, Prein, Delikaya, Haun, &

113 Gagarina, 2022). Similarly to the TANGO, the task was presented as an interactive web

114 application (see Figure 5; live demo <https://ccp-odc.eva.mpg.de/orev-demo/>; source code

115 <https://github.com/ccp-eva/orev-demo>).

116 Each trial presented four pictures: one target word alongside three distractors (1

117 phonological, 1 semantic, 1 unrelated distractor). A verbal prompt asked children to select

118 one of the pictures (prompt: “Zeige mir [target word]”; engl.: “Show me [target word]”).

119 The positioning of the target word and the three distractor types was counterbalanced

120 across the four display positions (upper/lower and left/right corners). Children responded

121 by clicking on one of the four pictures. The outcome measure consisted of the proportion of

122 correct responses.

123 **Procedure.** Families received a personalized study link and completed the study at

124 any location or time they wanted. Caregivers were asked to only provide technical support

125 but no further help. Webcam recordings were done whenever consented and technically

126 possible. We analyzed data of 20 trials per child. Children received the trials in two

127 different orders. Performance across the two study orders did not differ. Hence, we decided

128 to combine the data.

129 **Analysis.** To estimate the effect of gaze understanding on receptive vocabulary, we

130 fitted a GLMM predicting the task performance by the gaze understanding score (scaled),

131 controlling for age: `cor_clt | trials(n_clt) ~ age_centered +`

132 `mean_gaze_centered`). We modelled the number of participant’s correct responses using a

133 `Binomial` distribution.

134 **Results.**

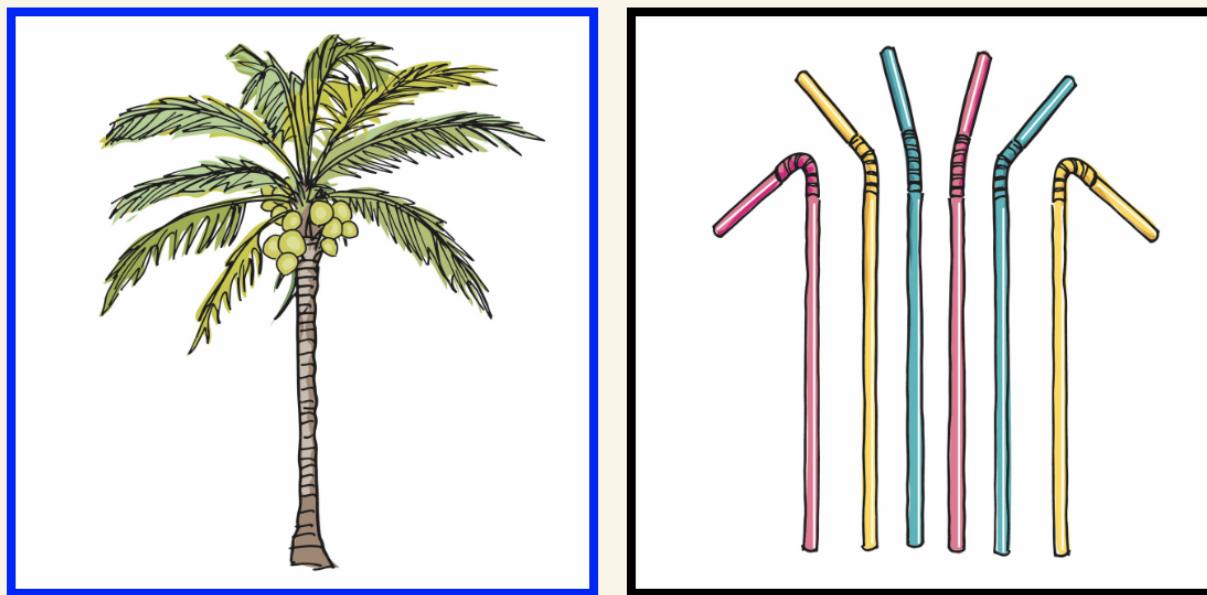
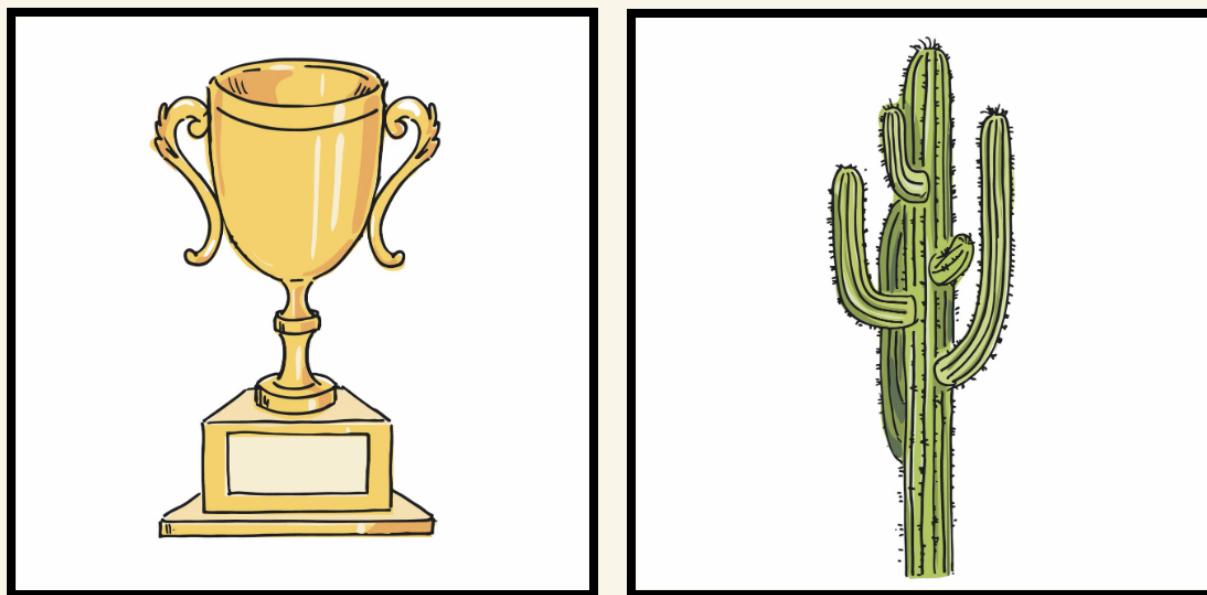
**WEITER**

Figure 5. Setup of the oREV. On each trial, participants heard a word and were asked to select the corresponding picture. Verbal prompts could be replayed by pressing the loudspeaker button.

Predictor	WAIC	SE_WAIC	Weight	ELPD_DIFF	SE_ELPD_DIFF
Age of childcare entry	571.66	28.13	0.73	0.00	0.00
Null model	575.85	30.39	0.09	-2.09	4.49
Sibling variety score (Peterson, 2000)	577.91	29.85	0.03	-3.12	4.64
# Children in household aged 0-18	578.01	29.70	0.03	-3.18	4.72
Peer exposure index	578.04	30.23	0.03	-3.19	4.70
Sibling variety score (Cassidy et al., 2005)	578.36	29.88	0.03	-3.35	4.65
# Household members	578.70	30.35	0.02	-3.52	4.50
# Children in household aged 1-12	578.93	30.04	0.02	-3.64	4.86
Average hours spent in childcare per day	579.03	30.69	0.02	-3.69	4.57

135

Additional information for the adult sample

Current Country of Residence	studytype	n
United Kingdom	vali	47
South Africa	vali	8
Portugal	vali	6
Poland	vali	5
Ireland	vali	5
Italy	vali	4
Spain	vali	4
Netherlands	vali	4
Canada	vali	2
Australia	vali	2
France	vali	2
Austria	vali	2
Norway	vali	1
DATA EXPIRED	vali	1
Finland	vali	1
Greece	vali	1
Germany	vali	1

(continued)

Current Country of Residence	studytype	n
United States	vali	1
Mexico	vali	1
South Africa	reli	48
United Kingdom	reli	19
United States	reli	14
Portugal	reli	10
Italy	reli	7
Canada	reli	6
Spain	reli	5
Ireland	reli	5
Poland	reli	4
Greece	reli	4
Netherlands	reli	3
CONSENT REVOKED	reli	3
Germany	reli	3
Australia	reli	2
France	reli	2
Czech Republic	reli	2
Hungary	reli	2
Chile	reli	1
Iceland	reli	1
New Zealand	reli	1

136 Recruitment

137 We recruited participants using the online participant recruitment service *Prolific*
138 from the University of Oxford. *Prolific*'s subject pool consists of a mostly European and
139 US-American sample although subjects from all over the world are included. The
140 recruitment platform realises ethical payment of participants, which requires researchers to
141 pay participants a fixed minimum wage of £5.00 (around US\$6.50 or €6.00) per hour. We
142 decided to pay all participants the same fixed fee which was in relation to the estimated
143 average time taken to complete the task. *Prolific* distributed our study link to potential
144 participants, while the hosting of the online study was done by local servers in the Max
145 Planck Institute for Evolutionary Anthropology, Leipzig. Therefore, study data was saved
146 only on our internal servers, while *Prolific* provided demographic information of the
147 participants. Participants' *Prolific* ID was forwarded to our study website using URL
148 parameters. This way, we could match participant demographic data to our study data.
149 The same technique was used to confirm study completion: we redirected participants from
150 our study website back to the *Prolific* website using URL parameters. We used *Prolific*'s
151 inbuilt prescreening filter to include only participants who were fluent in English and could
152 therefore properly understand our written and oral study instructions.

153 Study 1 - Validation hedge version

154 The aim of Study 1 was to validate the hedge version of our gaze understanding task.

155 The pre-registration can be found here: <https://osf.io/r3bhn>. We recruited participants
156 online by advertising the study on *Prolific*.

157 50 adults participated in the study. One additional subject returned their submission,

158 i.e., decided to leave the study early or withdrew their submission after study completion.

159 Data collection took place in May 2021. Participants were compensated with £1.25 for
160 completing the study. We estimated an average completion time of 6 minutes, resulting in

161 an estimated hourly rate of £10.00. On average, participants took 05:56min to complete
162 the study. Participants were required to complete the study on a tablet or desktop.
163 Participation on mobile devices was disabled since the display would be too small and
164 would harm click precision. It was indicated that the study required audio sound.

165 We stored *Prolific*'s internal demographic information, while not asking for additional
166 personal information.

167 **Study 2 - Validation box version**

168 As in study 1, we recruited participants on *Prolific*, and employed the same
169 methodology. However, this time we focussed on validating the box version of the task in
170 an adult sample. Participants were presented with eight boxes in which the target could
171 land. 50 adults participated in the study. One additional subject returned their
172 submission, i.e., decided to leave the study early or withdrew their submission after study
173 completion. Data collection took place in June 2021. Participants were compensated with
174 £1.00 for completing the study. We estimated an average completion time of 6 minutes,
175 resulting in an estimated hourly rate of £10.00. On average, participants took 04:43min to
176 complete the study.

177 **Study 3 - Reliability hedge version**

178 In study 3 and 4, we assessed the test-retest reliability of our gaze understanding task
179 in an adult sample. The pre-registration can be found here: <https://osf.io/nu62m>. We
180 tested the same participants twice with a delay of two weeks. The testing conditions were
181 as specified in Study 1 and 2. However, the target locations as well as the succession of
182 animals and target colors was randomized once. Each participant then received the same
183 fixed randomized order of target location, animal, and target color. Participants received
184 30 test trials without voice-over description, so that each of the ten bins occurred exactly

185 three times.

186 In addition to the aforementioned prescreening settings, we used a whitelist. *Prolific*
187 has a so-called *custom allowlist prescreening filter* where one can enter the *Prolific* IDs of
188 participants who completed a previous study. Only these subjects are then invited to
189 participate in a study. This way, repeated measurements can be implemented, collecting
190 data from the same subjects at different points in time.

191 In a first round, 60 participants took part on the first testday. Additional two
192 subjects returned their submission, i.e., decided to leave the study early or withdrew their
193 submission after study completion. One additional participant timed out, i.e., did not
194 finish the survey within the allowed maximum time. The maximum time is calculated by
195 *Prolific*, based on the estimated average completion time. For this study, the maximum
196 time amounted to 41 minutes. For the first testday, participants were compensated with
197 £1.25. We estimated an average completion time of 9 minutes, resulting in an estimated
198 hourly rate of £8.33. On average, participants took 07:11min to complete the first part.

199 Of the 60 participants that completed testday 1, 41 subjects finished testday 2. One
200 additional participant timed out, i.e., did not finish the survey within the allowed
201 maximum time. Participants were compensated with £1.50 for completing the second part
202 of the study. We estimated an average completion time of 9 minutes, resulting in an
203 estimated hourly rate of £10. On average, participants took 06:36min to complete the
204 second part of the study.

205 Since we aimed for a minimum sample size of 60 subjects participating on both
206 testdays, we reran the first testday with additional 50 participants. Additional seven
207 subjects returned their submission, i.e., decided to leave the study early or withdrew their
208 submission after study completion. Two additional participants timed out, i.e., did not
209 finish the survey within the allowed maximum time. Again, participants were compensated
210 with £1.25 for completing the first part of the study (estimated average completion time 9

211 minutes, estimated hourly rate of £8.33). On average, participants took 06:51min to
212 complete the first part.

213 Of the additional 50 participants that completed testday 1, 29 subjects finished
214 testday 2. Again, participants were compensated with £1.50 for completing the second
215 part of the study (estimated average completion time 9 minutes, estimated hourly rate of
216 £10). On average, participants took 06:26min to complete the second part of the study.

217 **Study 4 - Reliability box version**

218 As in study 3, we recruited participants on *Prolific*, and employed the same
219 methodology. However, this time participants were presented with the box version of the
220 task. Participants received 32 test trials without voice-over description, so that each of the
221 eight boxes occurred exactly four times. As in study 2, we employed eight boxes in which
222 the target could land.

223 In a first round, 60 participants took part on the first testday. Additional five
224 subjects returned their submission, i.e., decided to leave the study early or withdrew their
225 submission after study completion. For the first testday, participants were compensated
226 with £1.25. We estimated an average completion time of 9 minutes, resulting in an
227 estimated hourly rate of £8.33. On average, participants took 07:33min to complete the
228 first part.

229 Of the 60 participants that completed testday 1, 41 subjects finished testday 2.
230 Participants were compensated with £1.50 for completing the second part of the study. We
231 estimated an average completion time of 9 minutes, resulting in an estimated hourly rate of
232 £10. On average, participants took 07:50min to complete the second part of the study.

233 Since we aimed for a minimum sample size of 60 subjects participating on both
234 testdays, we reran the first testday with additional 50 participants. Additional eight
235 subjects returned their submission, i.e., decided to leave the study early or withdrew their

236 submission after study completion. One additional participant timed out, i.e., did not
237 finish the survey within the allowed maximum time. Again, participants were compensated
238 with £1.25 for completing the first part of the study (estimated average completion time 9
239 minutes, estimated hourly rate of £8.33). On average, participants took 07:37min to
240 complete the first part.

241 Of the additional 50 participants that completed testday 1, 28 subjects finished
242 testday 2. Additional three subjects returned their submission, i.e., decided to leave the
243 study early or withdrew their submission after study completion. One additional
244 participant timed out, i.e., did not finish the survey within the allowed maximum time.
245 Again, participants were compensated with £1.50 for completing the second part of the
246 study (estimated average completion time 9 minutes, estimated hourly rate of £10). On
247 average, participants took 06:30min to complete the second part of the study.

248

Instructions and voice-over descriptions

249 This is the content of our audio recordings that were played as instructions and
 250 during voice-over trials.

Timeline	German	English	Filename
welcome	Hallo! Schön, dass du da bist. Wir spielen jetzt das Ballon-Spiel! Siehst du die Tiere auf dem Bild da? Wir möchten gleich zusammen mit den Tieren mit einem Ballon spielen. Was genau passiert, erklären wir dir jetzt ganz in Ruhe.	Hello! Great that you're here. We'll now play a balloon game. Can you see the animals in the picture over there? We want to play together with the animals using the balloon. We'll now talk you through exactly what will happen.	welcome.mp3

touch	Schau mal, da steht ein Tier im Fenster. Und siehst du den Ballon da? Der Ballon fällt immer runter und landet auf dem Boden. Und du musst ihn dann finden. Das Tier hilft Dir und schaut immer den Ballon an.	Look, an animal is standing in the window. And can you see the balloon over there? The balloon always falls down and lands on the ground. And you have to find it! The animal helps you and always looks at the balloon.	touch-1.mp3
	Wo ist der Ballon? Drück auf den Ballon!	Where is the balloon? Click on the balloon!	prompt-touch- long.mp3

fam - HEDGE	Klasse, das war super! Jetzt spielen wir weiter. Siehst du wieder das Tier und den Ballon da? Der Ballon fällt wieder runter. Diesmal fällt er hinter eine Hecke. Du musst ihn wieder finden. Das Tier hilft dir und schaut immer den Ballon an.	Perfect, that was great! Now, we'll continue playing. Can you see the animal and the balloon again? The balloon will fall down again. This time, it will fall behind a hedge. And you have to find it! The animal helps you and looks at the balloon.	fam-hedge-1.mp3
	Wo ist der Ballon? Drücke auf die Hecke - wo der Ballon ist.	Where is the balloon? On the hedge, click where the balloon is.	prompt-hedge-long.mp3

fam - BOX	Klasse, das war super! Jetzt spielen wir weiter. Siehst du wieder das Tier und den Ballon da? Der Ballon fällt wieder runter. Diesmal fällt er in eine Kiste. Du musst ihn wieder finden. Das Tier hilft dir und schaut immer den Ballon an.	Perfect, that was great! Now, we'll continue playing. Can you see the animal and the balloon again? The balloon falls down again. This time, it falls into a box. And you have to find it!	fam-box-1.mp3
	Wo ist der Ballon? Drücke auf die Kiste mit dem Ballon.	Where is the balloon? Click on the box with the balloon.	prompt-box-long.mp3
test - HEDGE	Klasse , das hast du toll gemacht! Nun spielen wir weiter. Da sind wieder der Ballon, das Tier und die Hecke. Die Hecke wächst jetzt hoch.	Nice, good job! Now, we'll continue playing. There is the balloon, the animal and the hedge. The hedge is growing a bit now.	test-hedge-1.mp3

	Der Ballon ist nun hinter der Hecke. Du kannst das nicht sehen - das Tier aber! Jetzt fällt der Ballon auf den Boden und du musst ihn wieder finden. Denk dran - das Tier schaut immer den Ballon an.	The balloon is behind the hedge now. You can't see it - but the animal can! The balloon falls to the ground and you have to find it. Remember - the animal always looks at the balloon!	test-hedge-2.mp3
	Dann schrumpft die Hecke. Drücke auf die Hecke - wo der Ballon ist.	Now, the hedge is shrinking. On the hedge, click where the balloon is.	test-hedge-3.mp3
test - BOX	Klasse , das hast du toll gemacht! Nun spielen wir weiter. Da sind wieder der Ballon, das Tier und die Kisten. Jetzt wächst eine Hecke hoch.	Nice, good job! Now, we'll continue playing. There is the balloon and the animal. Now, a hedge is growing.	test-box-1.mp3

	Der Ballon ist nun hinter der Hecke. Du kannst das nicht sehen - das Tier aber! Jetzt fällt der Ballon in eine Kiste und du musst ihn wieder finden. Denk dran - das Tier schaut immer den Ballon an.	The balloon is behind the hedge now. You can't see it - but the animal can! The balloon falls into a box and you have to find it. Remember - the animal always looks at the balloon!	test-box-2.mp3
	Dann schrumpft die Hecke. Drücke auf die Kiste mit dem Ballon.	Now, the hedge is shrinking. Click on the box with the balloon.	test-box-3.mp3
goodbye	Geschafft! Die Tiere sind schon ganz glücklich vom Spielen! Vielen Dank für deine Hilfe! Bis zum nächsten Mal und liebe Grüße vom Schwein, Affen und Schaf	The animals are super happy after playing. Thanks a lot for your help! See you soon and goodbye from the pig, monkey and sheep	goodbye.mp3
general prompt	Wo ist der Ballon?	Where is the balloon?	prompt-general.mp3

touch - no response	Drück auf den Ballon!	Click on the balloon!	prompt-touch.mp3
hedge - no response	Drücke auf die Hecke - wo der Ballon ist!	On the hedge, click where the balloon is!	prompt-hedge.mp3
box - no response	Drücke auf die Kiste mit dem Ballon!	Click on the box with the balloon!	prompt-box.mp3
landing sound of balloon	-	-	balloon-lands.mp3
sound of blinking eyes	-	-	blink.mp3
sound for target click	-	-	positive-feedback.mp3

References

- Bohn, M., Prein, J., Delikaya, B., Haun, D. B. M., & Gagarina, N. (2022). *oREV: An Item Response Theory based open receptive vocabulary task for 3 to 8-year-old children.*
- Cassidy, K. W., Fineberg, D. S., Brown, K., & Perkins, A. (2005). Theory of Mind May Be Contagious, but You Don't Catch It from Your Twin. *Child Development*, 76(1), 97–106. Retrieved from <https://www.jstor.org/stable/3696711>
- McElreath, R. (2020). *Statistical rethinking: A Bayesian Course with Examples in R and Stan* (Second). Chapman and Hall/CRC.
- Peterson. (2000). Kindred spirits: Influences of siblings' perspectives on theory of mind. *Cognitive Development*, 15(4), 435–455.
[https://doi.org/10.1016/S0885-2014\(01\)00040-5](https://doi.org/10.1016/S0885-2014(01)00040-5)