1

- Measuring individual differences in the understanding of gaze cues across the lifespan
- Julia Prein<sup>1</sup>, Manuel Bohn<sup>1</sup>, Luke Maurits<sup>1</sup>, Steven Kalinke<sup>1</sup>, & Daniel M. Haun<sup>1</sup>
- $^{\rm 1}$  Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary
- 4 Anthropology, Leipzig, Germany

## Author Note

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

5

Abstract

There must be an abstract of no more than 250 words. One or two sentences providing a basic introduction to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "here we show" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

24 Keywords: social cognition, individual differences, gaze cues, psychometrics

Word count: X

Measuring individual differences in the understanding of gaze cues across the lifespan

# - POSSIBLE ROUGH STRUCTURE -

28	intro
29	individual differences in dev psy
30	• reliable tasks: need variation, more trials
31	exisiting tasks for social cognition
32	• wellman
33	current goal
34	• standardized, easy to use, continuous
35	methods (let's see whether we want this generic heading)
36	design of our task
37 38	$\bullet$ training trials (touch & fam together, click on visible balloon / visible target flight, voice over trials)
39	• test trials
40	• flexible: two versions, can also use discrete
41	• face value
42	• stimulus timing

# implementation / development

- JS, HTML, CSS
- parcel?!
- SVG: scalable
- webapp: portable across devices, flexible, no hard system requirements
- response collection: file format, variables saved (click responses)

## 49 data processing

- workflow for collecting data
- pipeline? downloading server data into r, coverting coordinate systems

### does our task induce variation?

- 53 participants: kiga & prolific vali
- 54 procedure
- 55 results
- for hedge & box
- sanity check: developmental trajectory
- remote testing: online child sample as side note

## do we capture variation reliably?

- 60 participants: relikiga & prolific reli
- 61 results:

59

- internal consistency?
- test-retest

71

64	exploratory: external validity
65	participants: subsample kiga & relikiga with questionnaire data
66	results:
67	• peer exposure (box & hedge combined?)
68	discussion
69	limitations
70	future development / extending the task

conclusion

### 72 Introduction

- Idea for an opener :)
- Developmental psychology is facing a dilemma: many research questions are
  questions about individual differences, yet, there is a lack of tasks to reliably measure these
- <sup>76</sup> individual differences. For example . . .
- individual differences in developmental psychology
  - reliable tasks, variation needed, more trials
- existing tasks for social cognition: wellman
  - goal of the current project

# Design of our balloon finding task

#### 82 Stimuli

78

80

81

- Our newly implemented balloon finding task displays cartoon-like stimuli that are
  engaging for children and adults alike. An animal character (i.e., agent; sheep, monkey, or
  pig) is placed centrally in a window. A balloon (i.e., target; blue, green, yellow, or red) is
  located in front of them. The target then falls to the ground. At all times, the agent's gaze
  tracks the movement of the target. That is, the pupils and iris of the agent move in a way
  that their center aligns with the center of the target. While the distance of the target's
  flight depends on the final location, the target moves at a constant speed. Participants are
  then asked to locate the target's position on the screen.
- To keep participants engaged and interested, the presentation of events is
  accompanied by cartoon-like effects. Each trial starts with an attention-getter: an
  eye-blinking sound plays while the pupils and iris of the agent enlarge (increase to 130%)
  and change in opacity (decrease to 75%) for 0.3 sec. The landing of the target is
  accompanied by a tapping sound. Once the target landed, the instructor's voice asks

"Where is the balloon?" After the response is registered, a short plop sound plays and a small orange circle confirms the participants' location choice. If no response got registered within 5 secs after the target landed, an audio prompt reminds the participant to respond.

#### 99 Trials

Before participants advance to the test trials, we present four training trials. In the 100 first training trial, participants have full visual access to the target flight and the target's 101 end location and are simply asked to click on the visible balloon. In the second and third 102 training trials, participants have partial visual access: they witness the target flight but 103 cannot see the target's end location. They are then asked to click on the invisible balloon, 104 i.e., the location where they saw the target land. In test trials, participants have no visual 105 access to the target flight nor the end location. Participants are expected to use the agent's 106 gaze as a cue to locate the target. The first trial of each type comprises a voice-over 107 description of the presented events. The audio descriptions explicitly state that the agent is 108 always looking at the target (see Appendix for audio script). After these four training and 109 audio guided trials, participants receive 15 test trials. The complete sequence of four 110 training trials and 15 test trials can be administered within 5-10 minutes of testing time.

### 112 Study versions

We designed two study versions which can be chosen according to the researchers'
need: there is a continuous *hedge* version and a discrete *box* version. Both versions use the
same first training trial and then differ in the consecutive training and test trials. In the *hedge* version, participants have to indicate their estimated target location directly on a
hedge (i.e., hedge version). Here, the dependent variable is imprecision, which is defined as
the absolute difference between the target's true x coordinate and the x coordinate of the
participant's click. In the *box* version, participants are asked to click on a box that hides
the target. Researchers have the choice of how many boxes are shown: one up to eight

boxes can be displayed as potential hiding locations. Here, we use a categorical outcome
(i.e., which box was clicked) to calculate the proportion of correct responses. Note that in
the test trials of both study versions, the target flight is covered by a hedge. In the *hedge*version, the hedge then shrinks to a minimum height required to cover the target's end
location. In the *box* version, the hedge shrinks completely. The boxes then hide the target's
final destination.



Figure 1. Study versions of the balloon finding task. (a) Screenshots from the continuous hedge version of the task. (b) Screenshots from the discrete box version of the task. (i) The agent stands in a window with the target in front of them. (ii) A hedge grows and covers the target. (iii) The target falls to a random location on the ground / into a box. The agent's eyes track the movement of the target.

#### Randomization

All agents and target colors appear equally often and are not repeated in more than two consecutive trials. The randomization of the target end location depends on the study version. In the *hedge* version, the full width of the screen is divided into ten bins. Exact coordinates within each bin are then randomly generated. In the *box* version, the target randomly lands in one of the boxes. As with agent and color choice, each bin/box occurs equally often and can only occur twice in a row.

## 34 Implementation

Our balloon finding task is presented as an interactive web-app. The task is portable across devices and web browsers and does not require any installation. A great advantage of online testing is that our testing procedure is standardized across participants. By using pre-recorded study instructions, no interaction with the experimenter is necessary during the study. The code is open-source (https://github.com/ccp-eva/gafo-demo) and a live demo version can be found under: https://ccp-odc.eva.mpg.de/gafo-demo/.

The web-app was programmed in JavaScript (ECMAScript 2015, i.e., ES6), HTML5, 141 CSS and PHP. For the web-app development, we chose the zero configuration bundler 142 Parcel. It enables a live server for debugging, makes imports easy, and minifies scripts in 143 order to save data storage. For the design of the landing pages, the CSS library Material 144 Components Web (TODO: insert citation) was used. For stimulus presentation, a scalable 145 vector graphic (SVG) composition was parsed. This way, the composition scales according 146 to the user's view port without loss of quality, while keeping the aspect ratio and relative 147 object positions constant. Furthermore, SVGs allow us to define all composite parts of the 148 scene (e.g., pupil of the agent) individually. This is needed for precisely calculating exact pupil and target locations and sizes. Additionally, it makes it easy to adjust the stimuli and, for example, add another agent to the scene. The GreenSock Animation Platform 151 (GSAP; TODO: insert citation) library was used to animate the movement of single SVG 152 elements. We use URL parameters to capture the participant's ID, language and study 153 version. 154

157

161

The web-app generates two file types: (1) a text file (.json) containing meta-data, 155 trial specifications and participants' click responses, and (2) a video file (.webm) of the 156 participant's webcam recording. For our samples described in this paper, we deployed the web-app on servers located in Leipzig, Germany. Data got automatically collected and 158 safely stored on these in-house servers. If no internet connection or server is available, 159 researchers could download a local version of the experiment which stores the generated 160 data automatically on the used device.



Figure 2. Infrastructure for online testing. With our new infrastructure, subjects aged 3 – 99+ can participate. Data collection is efficient and automatized and can take place anywhere: at home, in kindergartens or research institutes. Our task is presented as a website that works across devices. The scripts for the website and recorded data are stored on secure local servers.

## Data processing

164

165

166

167

168

workflow server to R, converting coordinate systems We used R [Version 4.1.2; R Core Team (2021) and the R-package papaja [Version 0.1.0.9997; Aust and Barth (2020) for all our analyses. For child samples, we uploaded safety copies of the already collected responses after the fourth test trial. In cases where children want to stop participation earlier, no responses get lost. All test trials without voice over description are included in our analyses.

# Does the balloon finding task induce variation?

# Participants

169

Our first aim was to assess whether our balloon finding task induces variation in a child and adult sample. The pre-registrations can be found here: https://osf.io/snju6 (child sample) and here: https://osf.io/r3bhn (adult sample). Participants were equally distributed across the two study versions.

For our child sample, we went to kindergartens in Leipzig and surroundings that 175 cooperate with the Max Planck Institute for Evolutionary Anthropology. Children in our 176 sample grow up in an industrialized, urban Central-European context. Information on 177 socioeconomic status was not formally recorded, although the majority of families come 178 from mixed, mainly mid to high socioeconomic backgrounds with high levels of parental 179 education. The child sample consisted of 120 children, including 40 3-year-olds (mean = 180 41.45 months, SD = 3.85, range = 36 - 47, 22 girls), 40 4-vear-olds (mean = 54.60 months), 181 SD = 3.10, range = 48 - 59, 19 girls), and 40 5-year-olds (mean = 66.95 months, SD =182 3.39, range = 60 - 71, 22 girls). Children received a small thank-you gift for their 183 participation in the study. 184

In addition, we recruited adult participants by advertising the study on *Prolific*.

Prolific is an online participant recruitment service from the University of Oxford with a predominantly European and US-american subject pool. Participants consisted of 100 English-speakers with an average age of 31.34 years (SD = 10.77, range = 18 - 63, 64 females). For completing the study, subjects were payed above the fixed minimum wage (in average £10.00 per hour). *Prolific* distributed our study link to potential participants, while the hosting of the online study was done by local servers in the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

#### 93 Procedure

All participants received 15 test trials that were displayed as described above. In the box version, we decided to adjust the task difficulty according to the participant's age:

children were presented with five boxes while adults were presented with eight boxes as

possible hiding location of the target.

#### 198 Results

200

201

- for hedge & box
  - sanity check: developmental trajectory
  - remote testing: online child sample as side note

To assess whether we would gain similar results in a remote, unsupervised testing 202 context with children, we recruited an additional online sample of 147 children, including 203 45 3-year-olds (mean = 42.62 months, SD = 3.35, range = 36 - 47, 14 girls), 47 4-year-olds204 (mean = 52.64 months, SD = 3.40, range = 48 - 59, 25 girls), and 55 5-year-olds (mean =205 65.11 months, SD = 3.77, range = 60 - 71, 27 girls). Families were recruited on a voluntary 206 basis via email from the institute internal database. In the beginning of the online study, 207 families were invited to enter "our virtual institute" and were welcomed by an introductory 208 video of the study leader, shortly describing the research background and further 200 procedure. Then, caregivers were informed about data security and were asked for their 210 informed consent. They were asked to enable the sound and seat their child centrally in 211 front of their device. Subsequently, a brief demographic questionnaire was displayed, asking for (1) the total number of household members, (2) the number of children, (3) age of the 213 other children, (4) whether the child was in day care, and if yes, (5) since when and (6) for how long on an average day. Before the study started, families were instructed how to 215 setup their webcam and enable the recording permissions. Study participation was video 216 recorded whenever possible in order to ensure that the answers were generated by the 217

children themselves. After completion, families received a little crafting / coloring sheet as 218 a small thank-you gift. Depending on the participant's device, the website automatically 219 presented the hedge or box version of the study. For families that used a tablet with 220 touchscreen, the hedge version was shown. Here, children could directly click on the 221 touchscreen themselves to indicate where the target is. For families that used a computer 222 without touchscreen, the website presented the box version of the task. We assumed that 223 younger children in our sample would not be acquainted with the usage of a computer 224 mouse. Therefore, we asked children to point to the screen. Caregivers were then asked to 225 act as the "digital finger" of their children and click on the indicated box. 226

# Can we capture variation reliably?

Our second aim was to assess whether the variation that we captured with our

# 228 Participants

227

229

237

238

239

females).

balloon finding task is reliable. The pre-registrations can be found here:

https://osf.io/xqm73 (child sample) and here: https://osf.io/nu62m (adult sample).

For our child sample, we again went to kindergartens in Leipzig and surroundings.

The child sample consisted of 87 children, including 19 3-year-olds (mean = 42.57 months,

SD = 3.32, range = 38 - 47, 9 girls), 38 4-year-olds (mean = 53.77 months, SD = 3.16,

range = 48 - 59, 19 girls), and 30 5-year-olds (mean = 66.10 months, SD = 3.43, range =

61 - 71, 16 girls).

The adult sample was again recruited over *Prolific* and consisted of 136

English-speakers with an average age of 25.74 years (SD = 8.11, range = 18 - 71, 87

#### 240 Procedure

The procedure was as in the first study, with the following differences. Participants 241 completed the study twice, with a delay of  $14 \pm 3$  days. The target locations as well as the 242 succession of animals and target colors was randomized once. Each participant then 243 received this fixed randomized order of target location, animal, and target color. The child 244 sample received 15 test trials. In the hedge version, each bin occurred once, making up ten 245 of the test trials. For the remaining five test trials, we repeated one out of two adjacent 246 bins (i.e., randomly chose between bin 1 & 2, bin 3 & 4, etc). In the box version, we 247 ensured that each of the five boxes occurred exactly three times. For the remaining training 248 trials, we repeated a fixed order of four random bins/boxes. Adults in the hedge version 240 received 30 test trials, each of the ten bin occurring exactly three times. Adults in the box version received 32 test trials with each of the eight boxes occurring exactly four times.

# 252 Results

253

255

- internal consistency
- test-retest

#### Exploring the external validity of our task

## 256 Participants

Our third aim was to assess whether the captured individual variation in gaze cue understanding relates to factors in children's real live social surroundings. For this exploratory analysis, we included all children of the aforementioned samples where families filled out a short demographic questionnaire. Our sample consisted of 130 children, including 39 3-year-olds (mean = 43.02 months, SD = 3.26, range = 37 - 47, 20 girls), 44 4-year-olds (mean = 54.43 months, SD = 2.77, range = 48 - 59, 26 girls), and 47 5-year-olds (mean = 66.19 months, SD = 3.56, range = 60 - 71, 23 girls).

#### Results

• peer exposure (hedge & box combined?)

Discussion

#### 267 Limitations

Future development / extending the task

269 Conclusion

270 Declarations

# Open practices statement

The web application (https://ccp-odc.eva.mpg.de/gafo-demo/) described here is open source (https://github.com/ccp-eva/gafo-demo). The datasets generated during and/or analysed during the current study are available in the [gazecues-methods] repository, (https://github.com/jprein/gazecues-methods). All experiments were preregistered (https://osf.io/zjhsc/).

## 277 Funding

282

This study was funded by the Max Planck Society for the Advancement of Science, a noncommercial, publicly financed scientific organization (no grant number). We thank all the children and parents who participated in the study.

## Conflicts of interest

The authors declare that they have no conflict of interest.

# 283 Ethics approval

# 284 Consent to participate

Informed consent was obtained from all individual participants included in the study or their legal guardians.

# 287 Consent for publication

Open access

# Authors' contributions

optional: please review the submission guidelines from the journal whether statements are mandatory

292	References
293	Aust, F., & Barth, M. (2020). papaja: Create APA manuscripts with R Markdown.
294	Retrieved from https://github.com/crsh/papaja
295	R Core Team. (2021). R: A language and environment for statistical computing.
296	Vienna, Austria: R Foundation for Statistical Computing. Retrieved from
297	https://www.R-project.org/

### Supplements

# 299 Adult sample

298

**Recruitment.** We recruited participants using the online participant recruitment 300 service Prolific from the University of Oxford. Prolific's subject pool consists of a mostly 301 European and US-american sample although subjects from all over the world are included. 302 The recruitment platform realises ethical payment of participants, which requires 303 researchers to pay participants a fixed minimum wage of £5.00 (around US\$6.50 or €6.00) per hour. We decided to pay all participants the same fixed fee which was in relation to the 305 estimated average time taken to complete the task. Prolific distributed our study link to 306 potential participants, while the hosting of the online study was done by local servers in the Max Planck Institute for Evolutionary Anthropology, Leipzig. Therefore, study data was saved only on our internal servers, while *Prolific* provided demographic information of 309 the participants. Participants' Prolific ID was forwarded to our study website using URL 310 parameters. This way, we could match participant demographic data to our study data. 311 The same technique was used to confirm study completion: we redirected participants from 312 our study website back to the *Prolific* website using URL parameters. We used *Prolific*'s 313 inbuilt prescreening filter to include only participants who were fluent in English and could 314 therefore properly understand our written and oral study instructions. 315

Study 1 - Validation hedge version. The aim of Study 1 was to validate the
hedge version of our balloon finding task. The pre-registration can be found here:
https://osf.io/r3bhn. We recruited participants online by advertising the study on *Prolific*.

50 adults participated in the study. One additional subject returned their submission,
i.e., decided to leave the study early or withdrew their submission after study completion.

Data collection took place in May 2021. Participants were compensated with £1.25 for
completing the study. We estimated an average completion time of 6 minutes, resulting in
an estimated hourly rate of £10.00. In average, participants took 05:56min to complete the

study. Participants were required to complete the study on a tablet or desktop.

Participation on mobile devices was disabled since the display would be too small and would harm click precision. It was indicated that the study required audio sound.

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

Study 2 - Validation box version. As in study 1, we recruited participants on 329 *Prolific*, and employed the same methodology. However, this time we focussed on 330 validating the box version of the task in an adult sample. Participants were presented with eight boxes in which the target could land. 50 adults participated in the study. One additional subject returned their submission, i.e., decided to leave the study early or 333 withdrew their submission after study completion. Data collection took place in June 2021. 334 Participants were compensated with £1.00 for completing the study. We estimated an 335 average completion time of 6 minutes, resulting in an estimated hourly rate of £10.00. In 336 average, participants took 04:43min to complete the study. 337

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

In addition to the beforementioned prescreening settings, we used a whitelist. *Prolific*has a so-called *custom allowlist prescreening filter* where one can enter the *Prolific* IDs of
participants who completed a previous study. Only these subjects are then invited to
participate in a study. This way, repeated measurements can be implemented, collecting

data from the same subjects at different points in time.

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

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

Since we aimed for a minimum sample size of 60 subjects participating on both 365 testdays, we reran the first testday with additional 50 participants. Additional seven 366 subjects returned their submission, i.e., decided to leave the study early or withdrew their 367 submission after study completion. Two additional participants timed out, i.e., did not 368 finish the survey within the allowed maximum time. Again, participants were compensated 369 with £1.25 for completing the first part of the study (estimated average completion time 9) 370 minutes, estimated hourly rate of £8.33). In average, participants took 06:51min to 371 complete the first part. 372

Of the additional 50 participants that completed testday 1, 29 subjects finished testday 2. Again, participants were compensated with £1.50 for completing the second part of the study (estimated average completion time 9 minutes, estimated hourly rate of

 $\pm 10$ ). In average, participants took 06:26min to complete the second part of the study.

Study 4 - Reliability box version. As in study 3, we recruited participants on

Prolific, and employed the same methodology. However, this time participants were

presented with the box version of the task. Participants received 32 test trials without

voice-over description, so that each of the eight boxes occurred exactly four times. As in

study 2, we employed eight boxes in which the target could land.

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

Of the 60 participants that completed testday 1, 41 subjects finished testday 2.

Participants were compensated with £1.50 for completing the second part of the study. We
estimated an average completion time of 9 minutes, resulting in an estimated hourly rate of
£10. In average, participants took 07:50min to complete the second part of the study.

Since we aimed for a minimum sample size of 60 subjects participating on both testdays, we reran the first testday with additional 50 participants. Additional eight subjects returned their submission, i.e., decided to leave the study early or withdrew their submission after study completion. One additional participant timed out, i.e., did not finish the survey within the allowed maximum time. Again, participants were compensated with £1.25 for completing the first part of the study (estimated average completion time 9 minutes, estimated hourly rate of £8.33). In average, participants took 07:37min to complete the first part.

Of the additional 50 participants that completed testday 1, 28 subjects finished testday 2. Additional three subjects returned their submission, i.e., decided to leave the

study early or withdrew their submission after study completion. One additional participant timed out, i.e., did not finish the survey within the allowed maximum time.

Again, participants were compensated with £1.50 for completing the second part of the study (estimated average completion time 9 minutes, estimated hourly rate of £10). In average, participants took 06:30min to complete the second part of the study.

# 7 Child sample

Study 1 - Validation Remote and in-person. The validation of our task in a in-person and remote child sample can be found here: https://osf.io/snju6. We chose to have at least 20 data points per cell (i.e. unique combination of data collection mode, study version, and age-group). Across the two data collection modes, a total of . . . children participated. Participants received a small gift as thank you for their participation in the study.

For our in-person supervised testing sample, we went to kindergartens in Leipzig and 414 surroundings that cooperate with the Max Planck Institute for Evolutionary Anthropology. 415 For our remote unsupervised testing sample, families were recruited on a voluntary basis 416 via email from the database of the Max Planck Institute for Evolutionary Anthropology. 417 Children in both sub samples live in Leipzig, Germany or surrounding areas and grow up 418 in an industrialized, urban Central-European context. Information on socioeconomic status 419 was not formally recorded, although the majority of families come from mixed, mainly mid to high socioeconomic backgrounds with high levels of parental education. Written 421 informed consent was obtained from at least one caregiver prior to testing. TODO: how to 422 phrase for kiga testing?

### Procedure Remote Testing

In the beginning of the online study, families were invited to enter "our virtual 425 institute" and were welcomed by an introductory video of the study leader, shortly 426 describing the research background and further procedure. Then, caregivers were informed 427 about data security and were asked for their informed consent. They were asked to enable 428 the sound and seat their child centrally in front of their device. Subsequently, a brief 429 demographic questionnaire was displayed, asking for (1) the total number of household 430 members, (2) the number of children, (3) age of the other children, (4) whether the child 431 was in day care, and if yes, (5) since when and (6) for how long on an average day. Before 432 the study started, families were instructed how to setup their webcam and enable the 433 recording permissions. Study participation was video recorded whenever possible in order to ensure that the answers were generated by the children themselves. Then, families were 435 guided through the online study with pre-recorded audio instructions. After completion, 436 families received a little crafting / coloring sheet as a small thank-you gift. 437

Depending on the participant's device, the website automatically presents the hedge 438 or box version of the study. For families that use a tablet with touchscreen, the hedge 439 version is shown. Here, children can directly click on the screen themselves to indicate 440 where the target is. For families that use a computer without touchscreen, the website 441 presents the box version of the task. We assumed that younger children in our sample 442 would not be acquainted with the usage of a computer mouse. Therefore, we asked children 443 to point to the screen. Caregivers were then asked to act as the "digital finger" of their 444 children and click on the indicated box. In order to facilitate the translation of children's 445 pointing and caregivers' clicking, we decided to implement this categorical version of the task. 447

Study 2 - Reliability in-person. https://osf.io/xqm73

448

449

450

# Instructions and voice over descriptions

This is the content of our voice recordings. Children listen to this during the game.

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

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

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

6	fam - BOX	Klasse, das war	Perfect, that	fam-box-1.mp3
		super! Jetzt	was great! Now,	
		spielen wir	we'll continue	
		weiter. Siehst	playing. Can	
		du wieder das	you see the	
		Tier und den	animal and the	
		Ballon da? Der	balloon again?	
		Ballon fällt	The balloon	
		wieder runter.	falls down	
		Diesmal fällt er	again. This	
		in eine Kiste.	time, it falls	
		Du musst ihn	into a box. And	
		wieder finden.	you have to find	
		Das Tier hilft	it! The animal	
		dir und schaut	helps you and	
		immer den	looks at the	
		Ballon an.	balloon.	
7		Wo ist der	Where is the	prompt-box-
		Ballon? Drücke	balloon? Click	long.mp3
		auf die Kiste	on the box with	
		mit dem Ballon.	the balloon.	

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

10		Dann schrumpft	Now, the hedge	test-hedge-
		die Hecke.	is shrinking. On	3.mp3
		Drücke auf die	the hedge, click	
		Hecke - wo der	where the	
		Ballon ist.	balloon is.	
11	test - BOX	Klasse , das	Nice, good job!	test-box-1.mp3
		hast du toll	Now, we'll	
		gemacht! Nun	continue	
		spielen wir	playing. There	
		weiter. Da sind	is the balloon	
		wieder der	and the animal.	
		Ballon, das Tier	Now, a hedge is	
		und die Kisten.	growing.	
		Jetzt wächst		
		eine Hecke		
		hoch.		

12	Der Ballon ist	The balloon is	test-box-2.mp3
	nun hinter der	behind the	
	Hecke. Du	hedge now. You	
	kannst das nicht	can't see it -	
	sehen - das Tier	but the animal	
	aber! Jetzt fällt	can! The	
	der Ballon in	balloon falls	
	eine Kiste und	into a box and	
	du musst ihn	you have to find	
	wieder finden.	it. Remember -	
	Denk dran - das	the animal	
	Tier schaut	always looks at	
	immer den	the balloon!	
	Ballon an.		
13	Dann schrumpft	Now, the hedge	test-box-3.mp3
	die Hecke.	is shrinking.	
	Drücke auf die	Click on the	
	Kiste mit dem	box with the	
	Ballon.	balloon.	

14	goodbye	Geschafft! Die	The animals are	goodbye.mp3
		Tiere sind schon	super happy	
		ganz glücklich	after playing.	
		vom Spielen!	Thanks a lot for	
		Vielen Dank für	your help! See	
		deine Hilfe! Bis	you soon and	
		zum nächsten	goodbye from	
		Mal und liebe	the pig, monkey	
		Grüße vom	and sheep	
		Schwein, Affen		
		und Schaf		
15	general	Wo ist der	Where is the	prompt-
	prompt	Ballon?	balloon?	general.mp3
16	touch - no	Drück auf den	Click on the	prompt-
	response	Ballon!	balloon!	touch.mp3
17	hedge - no	Drücke auf die	On the hedge,	prompt-
	response	Hecke - wo der	click where the	hedge.mp3
		Ballon ist!	balloon is!	
18	box - no	Drücke auf die	Click on the	prompt-
	response	Kiste mit dem	box with the	box.mp3
		Ballon!	balloon!	
19	landing sound	-	-	balloon-
	of balloon			lands.mp3
20	sound of	-	-	blink.mp3
	blinking eyes			

21	sound for -	-	positive-
	target click		feedback.mp3