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

2   Julia Prein[1], Manuel Bohn[1], Steven Kalinke[1], & Daniel M. Haun[1]

3   [1] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary

4   Anthropology, Leipzig, Germany

Abstract

To explain and predict the behavior of agents, we use social cognition: we represent and reason about others' perspectives, knowledge, intentions, beliefs, and preferences. However, traditional measures of social cognition (e.g., false belief change-of-location tasks) often lack satisfactory psychometric properties: they are not designed to capture variation *between* children and rely on low trial numbers, dichotomous measures, and group averages. This has profound implications for what these studies can show. Poor measurement of social cognition on an individual level may conceal relations between different aspects of cognition and may obscure developmental change. To fully understand how social-cognitive abilities emerge and relate to each other, we need new tools that can reliably measure individual differences. We designed a gaze understanding task to study social cognition in young children and adults to approach this issue. We concentrate on an essential ability that is involved in many social-cognitive reasoning processes: gaze understanding – the ability to locate and use the attentional focus of an agent. Our interactive task works across devices and enables supervised and unsupervised, as well as in-person and remote testing. The implemented spatial layout allows for discrete and continuous measures of participants' click imprecision and is easily adaptable to different study requirements. Here we show that our task induces inter-individual differences in a child (N = XXX) and an adult (N = XXX) sample. Our two study versions and data collection modes yield comparable results that show substantial developmental gains: the older children are, the more accurately they locate the target. High internal consistency and test-retest reliability estimates underline that the captured variation is systematic. This work shows a promising way forward in the study of individual differences in social cognition and will help us explore the in(ter)dependence of our core social-cognitive processes in greater detail.

Keywords: social cognition, individual differences, gaze cues, cognitive development

Word count: X

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

Introduction

Social cognition – representing and reasoning about an agent's perspectives, knowledge states, intentions, beliefs, and preferences to explain and predict their behavior – is among the most-studied phenomena in developmental research. In recent decades, much progress has been made in determining the average age at which a specific social-cognitive ability emerges in development (Gopnik & Slaughter, 1991; Peterson, Wellman, & Slaughter, 2012; Rakoczy, 2022; Wellman, Cross, & Watson, 2001; Wellman & Liu, 2004). Yet, there are always individual differences. Identifying variability in social-cognitive abilities and factors influencing their development is vital in theory building (e.g., to test causal predictions) and for designing interventions (Happé, Cook, & Bird, 2017; Kidd, Donnelly, & Christiansen, 2018; Lecce, Bianco, Devine, Hughes, & Banerjee, 2014; Mundy et al., 2007; Underwood, 1975).

Numerous studies have already examined individual differences in social cognition (for an overview, see Hughes & Devine, 2015; Slaughter, 2015). These individual differences studies often focus on the relationship between social-cognitive abilities and: (1) family influences, (2) other cognitive constructs, and (3) social behavioral outcome (for an overview, see Repacholi, 2003). Studies on social-cognitive abilities and family influences include the effect of parenting practices (for a review, see Pavarini, de Hollanda Souza, & Hawk, 2013), attachment quality (e.g., Astor et al., 2020), mental state talk (Gola, 2012; Hughes, Ensor, & Marks, 2011; Lecce et al., 2014), and family background as parental education, occupation, sibling interaction and childcare (Bulgarelli & Molina, 2016; Cutting & Dunn, 1999; Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991). Another group of individual differences studies focuses on the interplay of social and physical cognition (Herrmann, Hernández-Lloreda, Call, Hare, & Tomasello, 2010), executive functions (Benson, Sabbagh, Carlson, & Zelazo, 2012; Buttelmann, Kühn, & Zmyj, 2021;

Carlson & Moses, 2001; Carlson, Moses, & Claxton, 2004; Hughes & Ensor, 2007), and

language abilities (McEwen et al., 2007; Milligan, Astington, & Dack, 2007; Okumura,

Kanakogi, Kobayashi, & Itakura, 2017). Studies on social behavioral outcomes measured

the interplay of social cognition and prosociality (for a review, see Imuta, Henry, Slaughter,

Selcuk, & Ruffman, 2016; Walker, 2005), stereotypes and resource allocations (Rizzo &

Killen, 2018), and moral intentions (Sodian et al., 2016).

However, frequently, developmental psychologists are surprised to find minor or no

association between measures of social cognition that are thought to be theoretically

related – cross-sectionally and/or longitudinally (e.g., Sodian et al., 2016). This might be

because traditional measures of social cognition are not designed to capture variation

between children: they often rely on low trial numbers, small sample sizes, and

dichotomous measures. A recent review showed that many studies on social cognition

measures failed to report relevant psychometric properties at all (Beaudoin, Leblanc,

Gagner, & Beauchamp, 2020) or – when they did – showed mixed results on test-retest

reliability (Hughes et al., 2000; Mayes, Klin, Tercyak, Cicchetti, & Cohen, 1996).

To give an example: perhaps the most commonly applied prototypical measure for

social cognition is the change-of-location false belief task (Baron-Cohen, Leslie, & Frith,

1985; Wimmer & Perner, 1983). Here, children watch a short sequence of events (often

acted out or narrated by the experimenters). A doll called Sally puts her marble into a

basket. After Sally leaves the scene, a second doll named Anne takes the marble and moves

it into a box. Participants then get asked where Sally will look for her marble once she

returns. The outcome measures false belief understanding in a dichotomous way: children

pass the task if they take the protagonist's epistemic state into account and answer that she

will look into the basket. Many years of research utilizing these verbal change-of-location

tasks suggest that children develop belief-representing abilities at four to five years of age

(for a review, see Wellman et al., 2001). Several cross-cultural studies supported this

evidence (Barrett et al., 2013; Callaghan et al., 2005; cf. Mayer & Träuble, 2015).

⁸⁸   However, from this age onwards, the change-of-location task shows ceiling effects and

⁸⁹ has very limited diagnostic value (Repacholi, 2003). Thus, this sort of task seems to be well

⁹⁰ suited to track group-level developmental trends, yet it fails to capture individual

⁹¹ differences (cf. "reliability paradox," Hedge, Powell, & Sumner, 2018). As Wellman (2012)

⁹² put it, "it's really only passing/failing one sort of understanding averaged across age"

⁹³ (p. 317). This has profound implications for what studies on individual differences using

⁹⁴ this task (or others) can show. Poor measurement of social cognition on an individual level

⁹⁵ is likely to conceal relations between different aspects of cognition and may obscure

⁹⁶ developmental change.

⁹⁷   Thus, developmental psychology faces a dilemma: many research questions are

⁹⁸ questions about individual differences, yet, there is a lack of tasks to measure these

⁹⁹ individual differences reliably. To capture the emergence of social-cognitive abilities and

¹⁰⁰ related social factors in greater precision and detail, we must consequently address the

¹⁰¹ methodological limitations of existing study designs (Hughes et al., 2011; Hughes &

¹⁰² Leekam, 2004).

¹⁰³   Schaafsma, Pfaff, Spunt, and Adolphs (2015) compiled a "wish-list" for new

¹⁰⁴ social-cognitive paradigms. They advocated for parametric – instead of dichotomous –

¹⁰⁵ measures covering proficiency as a range, avoiding floor and ceiling effects, and showing

¹⁰⁶ satisfactory test-retest reliability estimates (see also Beaudoin et al., 2020; Hughes &

¹⁰⁷ Devine, 2015). New tasks should induce variation across age groups, including older

¹⁰⁸ children and adults (Repacholi, 2003). Another goal in creating new tasks should be to

¹⁰⁹ focus on the "face value": measures should probe the underlying social-cognitive ability as

¹¹⁰ straight-forward and directly as possible. Keeping task demands minimal is also beneficial

¹¹¹ for using the paradigm in a variety of different cultural, clinical and demographic contexts

¹¹² (Molleman, Kurvers, & van den Bos, 2019). The task should serve as a proxy for behavior

¹¹³ as it appears in the real world and should be validated in relation to real-world experiences

¹¹⁴ (Repacholi, 2003).

A new measure of gaze understanding

Our goal was to design a new measure of social cognition that captures individual differences across age groups in a systematic and reliable way. We focused on a fundamental ability implicated in many social-cognitive reasoning processes: gaze understanding – the ability to locate and use the attentional focus of an agent. The first component of this ability is often termed gaze following – turning one's eyes in the same direction as the gaze of another agent – and has been studied intensively (Astor, Thiele, & Gredebäck, 2021; Byers-Heinlein et al., 2021; Coelho, George, Conty, Hugueville, & Tijus, 2006; Del Bianco, Falck-Ytter, Thorup, & Gredebäck, 2019; Frischen, Bayliss, & Tipper, 2007; Hernik & Broesch, 2019; Itakura & Tanaka, 1998; Lee, Eskritt, Symons, & Muir, 1998; Moore, 2008; Shepherd, 2010; Tomasello, Hare, Lehmann, & Call, 2007). Following an agent's gaze provides insights into their intentions, thoughts, and feelings by acting as a "front end ability" (Brooks & Meltzoff, 2005, p. 535). In our definition, gaze understanding goes one step further by including the *acting on the gaze-cued location* – therefore, using the available social information to guide one's behavior as needed in real-life conditions.

To address the psychometric shortcoming of earlier work, we implemented the following design features: First, we used a continuous measure which allowed us to capture fine grained individual differences at different ages. Second, we designed short trials that facilitate more than a dozen replicates per subject. The result are more precise individual-level estimates. Third, we systematically investigated the psychometric properties of the new task.

Designing this task required a new testing infrastructure. We designed the task as an interactive browser-based web app. This greatly increased the flexibility with which we could modify the stimuli on a trial-by-trial basis. Furthermore, because the task is largely self-contained, it is much more controlled and standardized. Most importantly perhaps, it makes the task portable: testing is possible in-person using tablets but also remotely via

the internet (no installation needed). As such, it provides a solid basis to study individual

differences in gaze understanding across ages at scale. We make the task and its source

code openly accessible for other researchers to use and modify.

Task design

Implementation

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`, `HTML5`, `CSS` and `PHP`. For stimulus

presentation, a scalable vector graphic (SVG) composition was parsed. This way, the

composition scales according to the user's view port without loss of quality, while keeping

the aspect ratio and relative object positions constant. Furthermore, SVGs allow us to

define all composite parts of the 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 web app

generates two file types: (1) a text file (.json) containing meta-data, trial specifications and

participants' click responses, and (2) a video file (.webm) of the participant's webcam

recording. These files can either be sent to a server or downloaded to the local device.

Stimuli

Our newly implemented task features an online game where children or adults are

asked to search for a balloon. The events proceed as follows (see Figure 1B and C). An

animated agent (a sheep, monkey, or pig) looks out of a window of a house. 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

165 target. While the distance of the target's flight depends on the final location, the target

166 moves at a constant speed. Participants are then asked to locate the target: they respond

167 by touching or clicking on the screen. Visual access to the target's true location is

168 manipulated by a hedge. Participants either have full, partial, or no visual access to the

169 true target location. When partial or no information about the target location is accessible,

170 participants are expected to use the agent's gaze as a cue.

171   To keep participants engaged and interested, the presentation of events is

172 accompanied by cartoon-like effects. Each trial starts with an attention-getter: an

173 eye-blinking sound plays while the pupils and iris of the agent enlarge (increase to 130%)

174 and change in opacity (decrease to 75%) for 0.3 sec. The landing of the target is

175 accompanied by a tapping sound. Once the target landed, the instructor's voice asks

176 "Where is the balloon?". For confirming the participant's click, a short plop sound plays

177 and a small orange circle appears at the location of choice. Participants do not receive

178 differential feedback so that learning effects are reduced and trials stay comparable across

179 the sample. If no response is registered within 5 secs after the target landed, an audio

180 prompt reminds the participant to respond.

181 **Trials**

182   Trials differ in the amount of visual access that participants have to the final target

183 position. Before the test trials start, participants complete four training trials during which

184 they familiarize themselves with clicking the screen. In the first training trial, participants

185 have full visual access to the target flight and the target's end location and are simply

186 asked to click on the visible balloon. In the second and third training trials, participants

187 have partial access: they witness the target flight but cannot see the target's end location.

188 They are then asked to click on the hidden balloon, i.e., the location where they saw the

189 target land. In test trials, participants have no visual access to the target flight or the end

190 location. Participants are expected to use the agent's gaze as a cue to locate the target.

191 The first trial of each type comprises a voice-over description of the presented events. The

192 audio descriptions explicitly state that the agent is always looking at the target (see

193 Appendix for audio script). After the four training trials, participants receive 15 test trials.

194 The complete sequence of four training trials and 15 test trials can be easily completed

195 within 5-10 minutes.

Study versions

197 We designed two study versions which differ in the final hiding place of the target

198 and, consequently, on the outcome measure: a *hedge version* (continuous) and a *box version*

199 (discrete). Both versions use the same first training trial and then differ in the consecutive

200 training and test trials. In the hedge version, participants have to indicate their estimated

201 target location directly on a hedge. Here, the dependent variable is imprecision, which is

202 defined as the absolute difference between the target center and the x coordinate of the

203 participant's click. In the box version, the target lands in a box and participants are asked

204 to click on the box that hides the target. Researchers have the choice how many boxes are

205 shown: one up to eight boxes can be displayed as potential hiding locations. Here, we use a

206 categorical outcome (i.e., which box was clicked) to calculate the proportion of correct

207 responses. Note that in the test trials of both versions, the target flight is covered by a

208 hedge. In the hedge version, the hedge then shrinks to a minimum height required to cover

209 the target's end location. In the box version, the hedge shrinks completely. The boxes then

210 hide the target's final destination (see Figure 1B and C).

Randomization

212 All agents and target colors appear equally often and are not repeated in more than

213 two consecutive trials. The randomization of the target end location depends on the study

214 version. In the hedge version, the full width of the screen is divided into ten bins. Exact

215 coordinates within each bin are then randomly generated. In the box version, the target
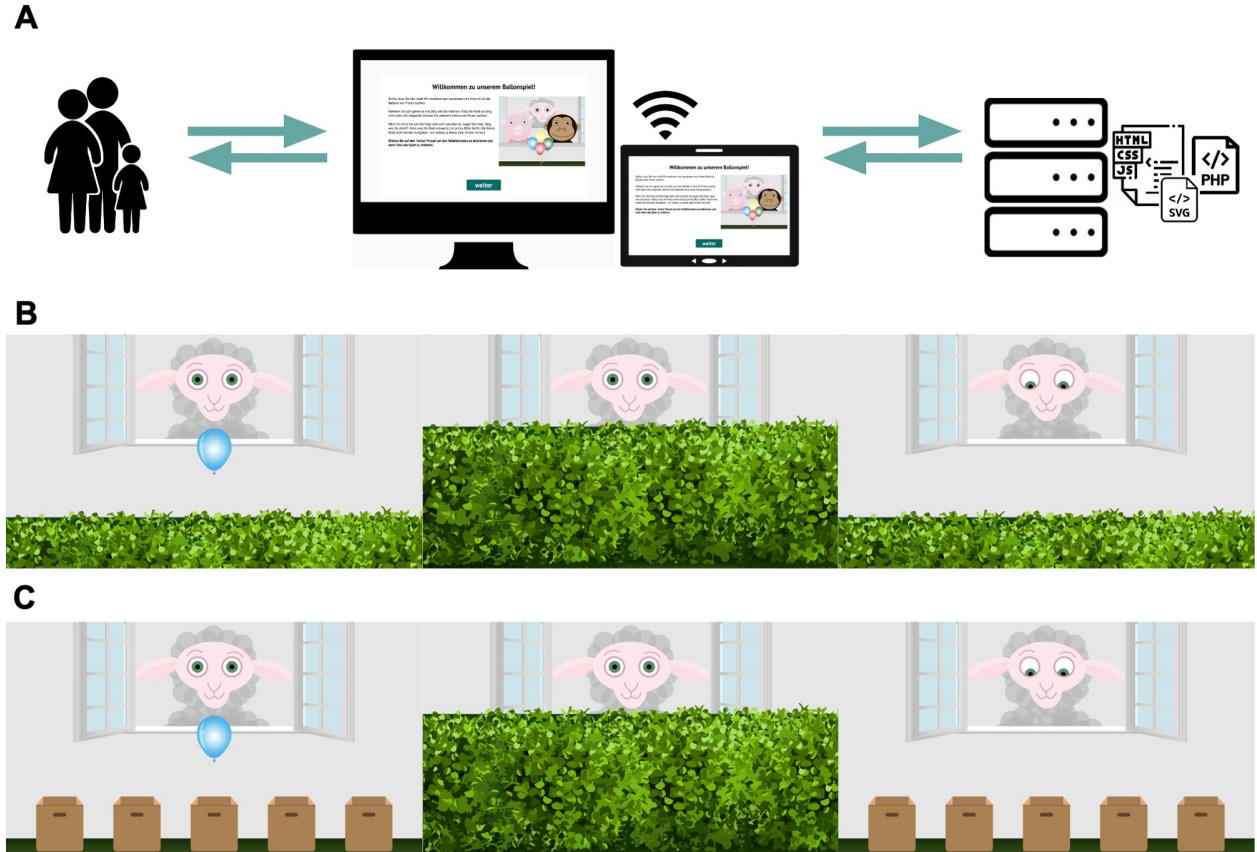
Figure 1. **Study setup**. (A) Infrastructure for online testing. (i) Subjects aged $3 - 99+$ can participate. Data collection can take place anywhere: online, in kindergartens or research labs. (ii) The task is presented as a website that works across devices. (iii) The scripts for the website and the recorded data are stored on secure local servers. (B) Hedge version (continuous) of the gaze understanding 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. The agent's eyes track the movement of the target. (C) Box version (discrete) of the gaze understanding task. Number of boxes (min. 1; max. 8) as potential hiding locations can be set according to the researcher's need.

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.

Individual differences

Our first aim was to assess whether our gaze understanding task induces
inter-individual variation in a child and adult sample. Furthermore, we were interested in
how the data collection mode (in-person vs. remote) influences responses.

Methods, sample size and analysis were pre-registered: https://osf.io/snju6 (child
sample) and https://osf.io/r3bhn (adult sample). Participants were equally distributed
across the two study versions. The study was approved by an internal ethics committee at
the Max Planck Institute for Evolutionary Anthropology. Data was collected between May
and October 2021.

Participants

We collected data from an in-person child sample, a remote child sample, and a
remote adult sample. In-person testing with children took place in kindergartens in
Leipzig, Germany. The in-person child sample consisted of 120 children, including 40
3-year-olds (mean = 41.45 months, SD = 3.85, range = 36 - 47, 22 girls), 40 4-year-olds
(mean = 54.60 months, SD = 3.10, range = 48 - 59, 19 girls), and 40 5-year-olds (mean =
66.95 months, SD = 3.39, range = 60 - 71, 22 girls).

For our remote child sample, we recruited families via an internal database of
children living in Leipzig, Germany, whose parents volunteered to participate in studies on
child development and who indicated interest in online studies. Families received an email
with a short study description and a personalized link, and received a reminder two weeks
later if they had not already participated in the study.

The remote child sample included 147 children, including 45 3-year-olds (mean =

42.62 months, SD = 3.35, range = 36 - 47, 14 girls), 47 4-year-olds (mean = 52.64 months,

SD = 3.40, range = 48 - 59, 25 girls), and 55 5-year-olds (mean = 65.11 months, SD =

3.77, range = 60 - 71, 27 girls). Children in our sample grow up in an industrialized, urban

Central-European context. Information on socioeconomic status was not formally recorded,

although the majority of families come from mixed, mainly mid to high socioeconomic

backgrounds with high levels of parental education.

Adults were recruited via *Prolific* (Palan & Schitter, 2018). *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).

Procedure

Children in our in-person sample were tested on a tablet in a quiet room in their

kindergarten. An experimenter guided the child through the study. Children in the remote

sample received a personalized link to the study website and families could participate at

any time or location they wanted. In the beginning of the online study, families were

invited to enter our "virtual institute" and were welcomed by an introductory video of the

study leader, shortly describing the research background and further procedure. Then,

caregivers were informed about data security and were asked for their informed consent.

They were asked to enable the sound and seat their child centrally in front of their device.

Before the study started, families were instructed how to setup their webcam and enable

the recording permissions. We stressed that caregivers should not help their children.

Study participation was video recorded whenever possible in order to ensure that the

answers were generated by the children themselves. Depending on the participant's device,

the website automatically presented the hedge or box version of the study. For families that

used a tablet with touchscreen, the hedge version was shown. Here, children could directly

click on the touchscreen themselves to indicate where the target is. For families that used a

computer without touchscreen, the website presented the box version of the task. We

assumed that younger children in our sample would not be acquainted with the usage of a

computer mouse. Therefore, we asked children to point to the screen, while caregivers were

asked to act as the "digital finger" of their children and click on the indicated box.

All participants received 15 test trials. In the box version, we decided to adjust the

task difficulty according to the sample: children were presented with five boxes while

adults were presented with eight boxes as possible target locations.

Analysis

All test trials without voice over description were included in our analyses. We ran all

analyses in R version 4.2.0 (2022-04-22) (R Core Team, 2022). Regression models were fit

as Bayesian generalized linear mixed models (GLMMs) with default priors for all analyses,

using the function `brm` from the package `brms` (Bürkner, 2017, 2018).

To estimate the developmental trajectory of gaze understanding and the effect of data

collection mode, we fit a GLMM predicting the task performance in each trial by age (in

months, z-transformed) and data collection mode (reference category: in-person

supervised). The model included random intercepts for each participant and each target

position, and a random slope for symmetric target position within participants (model

notation in R: `performance ~ age + datacollection + (symmetricPosition |`

`subjID) + (1 | targetPosition)`). Here, `targetPosition` refers to the exact bin/box of

the target, while `symmetricPosition` refers to the absolute distance from the stimulus

center (i.e., smaller value meaning more central target position). We expected that trials

could differ in their difficulty depending on the target centrality and that these these item

effects could vary between participants.

For the hedge version, performance was defined as the absolute click distance between the target center and the click X coordinate, scaled according to target widths, and modeled by a `lognormal` distribution. For the box version, the model predicted correct responses (0/1) using a `Bernoulli` distribution with a logit link function. We inspected the posterior distribution (mean and 95% Confidence Interval (CI)) for the age and data collection estimates.

Results

We found a strong developmental effect: with increasing age, participants got more and more accurate in locating the target. In the hedge version, children's click imprecision decreased with age, while, in the box version the proportion of correct responses increased (see Figure 2A and F). Most participants in the box version performed above chance level. By the end of their sixth year of life, children came close to the adult's proficiency level. Most importantly, however, we found substantial inter-individual variation across study versions and age groups. For example, some three-year-olds were more precise in their responses than some five-year-olds. Even though variation is smaller, we could even find inter-individual differences in the adult sample.

As Figure 2A and F show, our remotely collected child data resembled the data from the kindergarten sample. We found evidence that responses of children participating remotely were slightly more precise. This difference was mainly driven by the younger participants and especially prominent in the box version of the task. It is conceivable that caregivers were especially prone to influence the behavior of younger children. In the box version, caregivers might have had more opportunities to interfere since they carried out the clicking for their children.[1]

---

[1] In an exploratory analysis, we coded parental behavior and environmental factors during remote unsupervised testing. We focused on the subsample with the greatest performance difference between data collection modes: the three-year-olds in the box version of the task (n = 16). We reasoned that if parental interference cannot explain the greatest performance difference in our sample, the effects would be
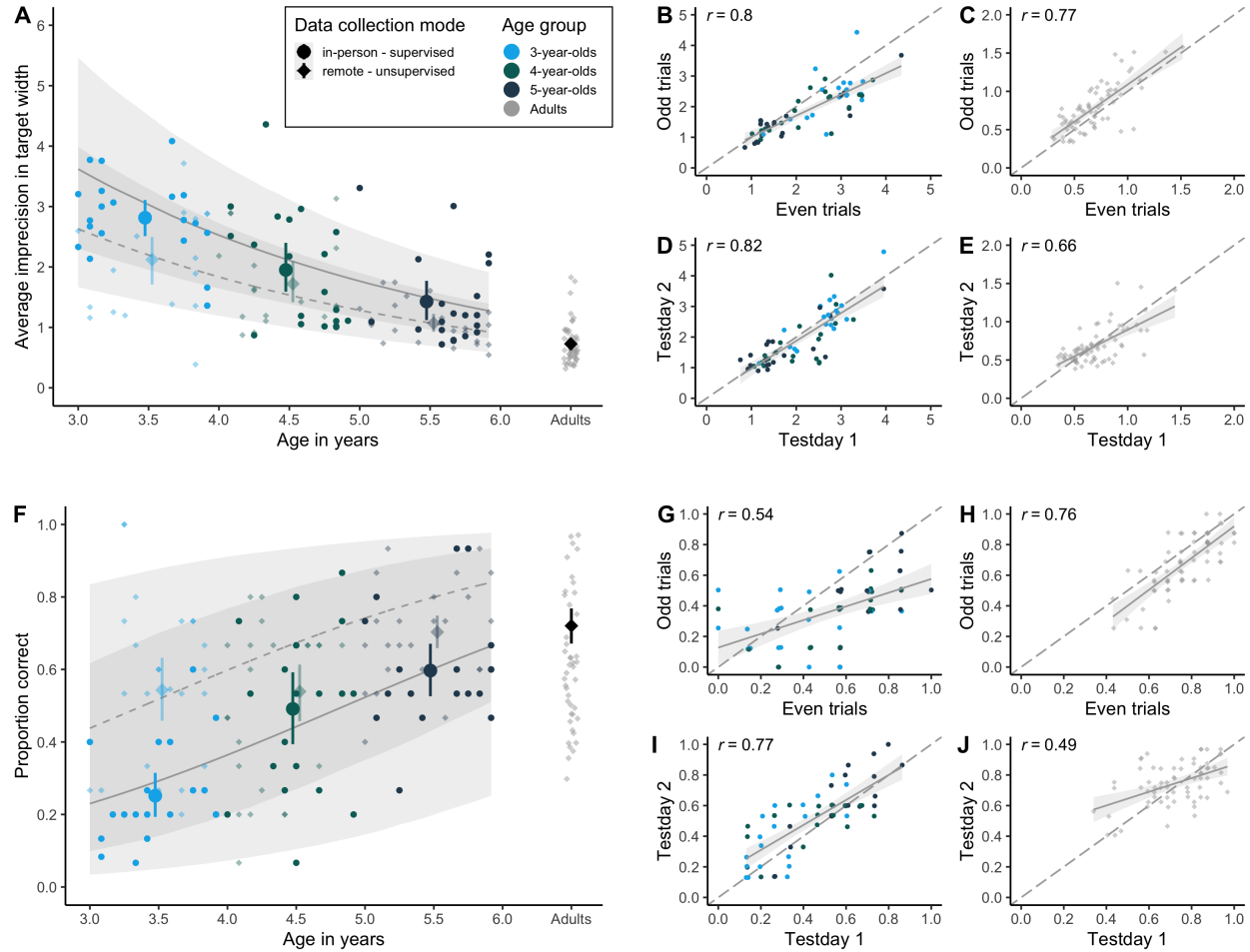
Figure 2. **Measuring inter-individual variation**. (A) Developmental trajectory in continuous hedge version. Performance is measured as imprecision, i.e., the absolute distance between the target's center and the participant's click (averaged across trials). The unit of imprecision is counted in the width of the target, i.e., a participant with an imprecision of 1 clicked in average one target width to the left or right of the true target center. (B) Internal consistency (odd-even split) in hedge child sample. (C) Internal consistency in hedge adult sample. (D) Test-retest reliability in hedge child sample. (E) Test-retest reliability in hedge adult sample. (F) Developmental trajectory in discrete box version. Performance is measured as the proportion of correct responses, i.e., how many times the participant clicked on the box that actually contained the target. Dotted black line shows level of performance expected by chance (for child sample 20%, i.e., 1 out of 5 boxes; for adult sample 12.5%, i.e., 1 out of 8 boxes). (G) Internal consistency (odd-even split) in box child sample. (H) Internal consistency in box adult sample. (I) Test-retest reliability in box child sample. (J) Test-retest reliability in box adult sample. For (A) and (F), regression lines show the predicted developmental trajectories (with 95% CrI) based on GLMMs, with the line type indicating the data collection mode. For (B-E) and (G-J), 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*. Large points with 95% CI (based on non-parametric bootstrap) represent performance means by age group (binned by year). Small points show the mean performance for each subject averaged across trials. Shape of data points represents data collection mode: opaque circles for in-person supervised data collection, translucent diamonds for remote unsupervised data collection. Color of data points denotes age group.

₃₁₄   Our GLMM analysis corroborated the visual inspection of the data: in the hedge

₃₁₅ version, the estimates for age (β = -0.33; 95% CI [-0.41; -0.24]) and data collection mode

₃₁₆ -0.32 (95% CI [-0.50; -0.14]) were negative and reliably different from zero. In the box

₃₁₇ version, the estimate of age (β =0.63 (95% CI [0.40; 0.87]) and the estimate of data

₃₁₈ collection mode (β = 1.12 (95% CI [0.68; 1.56]) were positive and reliably different from

₃₁₉ zero. Note that even though confidence intervals from the data collection estimates were

₃₂₀ wide, the effect was positive and reliably different from zero in a way that our remote

₃₂₁ sample performed more accurately than our in-person sample.

Discussion

₃₂₃   Our task induced inter-individual variation in both children and adults. We see

₃₂₄ substantial developmental gains: with increasing age, participants got more and more

₃₂₅ precise in locating the target. The five-year-olds reached a proficiency level close to the

₃₂₆ adults' level. For neither study version nor age group did we find any floor or ceiling

₃₂₇ effects. The presentation as a tablet game kept children interest and motivated throughout

₃₂₈ the 15 test trials. Furthermore, we found a comparable developmental trajectory for an

₃₂₉ unsupervised remote child sample. This illustrates the flexibility of the task design.

Internal consistency and retest reliability

₃₃₁   As a next step, we aimed at investigating whether the variation that we captured

₃₃₂ with our gaze understanding task is reliable. We assessed internal consistency (as splithalf

₃₃₃ reliability) and test-retest reliability. Data collection and analysis were pre-registered (can

₃₃₄ be found here: https://osf.io/xqm73 for the child sample, and https://osf.io/nu62m for the

₃₃₅ adult sample). Participants were equally distributed across the two study versions. The

―――――

negligible in the remaining sample. Based on our model comparison, we conclude that there is no clear
evidence of a stable effect of parental interference. See Supplements for further detail.

study was approved by an internal ethics committee at the Max Planck Institute for

Evolutionary Anthropology. Data was collected between July 2021 and June 2022.

Participants

Participants were recruited in the same way as in the previous study. The child

sample consisted of 120 children, including 41 3-year-olds (mean = 42.34 months, SD =

3.10, range = 37 - 47, 20 girls), 41 4-year-olds (mean = 53.76 months, SD = 3.15, range =

48 - 59, 21 girls), and 38 5-year-olds (mean = 66.05 months, SD = 3.40, range = 60 - 71, 19

girls).

The adult sample consisted of 136 English-speakers with an average age of 25.73

years (SD = 8.09, range = 18 - 71, 87 females).

Procedure

We applied the same procedure as in the first study, with the following differences.

Participants completed the study twice, with a delay of 14 ± 3 days. The target locations

as well as the succession of agents and target colors was randomized once and then held

constant across participants. The child sample received 15 test trials. In the hedge version,

each bin occurred once, making up ten of the test trials. For the remaining five test trials,

we repeated one out of two adjacent bins (i.e., randomly chose between bin 1 & 2, bin 3 &

4, etc). In the box version, we ensured that each of the five boxes occurred exactly three

times. For the remaining training trials, we repeated a fixed order of four random

bins/boxes. Adults in the hedge version 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.

Analysis

We assessed reliability in two ways. First, we focused on the internal consistency by calculating splithalf reliability coefficients. For each subject, trials were split into odd and even trials, performance was aggregated and then correlated using *Pearson* coefficients. For this, we used the data of the first test day. Performance was defined according to study version: in the hedge version, performance referred to the mean absolute difference between the target center and the click coordinate, scaled according to target widths; in the box version, we computed the mean proportion of correct choices. Pronk, Molenaar, Wiers, and Murre (2021) recently compared various methods for computing splithalf reliability that differ in how the trials are split into parts and whether they are combined with stratification by task design. To compare our traditional approach of a simple odd-even split, we additionally calculated reliability estimates using first-second, odd-even, permutated, and Monte Carlo splits without and with stratification by target position. First-second and odd-even splits belong to single sample methods, since each participant has a single pair of performance scores, while permutated (without replacement) and Monte Carlo (with replacement) splits make use of resampling. Analyses were run using the function `by_split` from the `splithalfr` package (Pronk et al., 2021).

Second, we assessed the test-retest reliability. We calculated performance scores (depending on study version as described above) for each participant in each test session and correlated them using *Pearson* correlation coefficients. Furthermore, for our child sample we report an age-corrected correlation between the two test days using a GLMM based approach (Rouder & Haaf, 2019). We fit trial-by-trial data with a fixed effect of age, a random intercept for each subject and a random slope for test day (model notation in `R`: `performance ~ age (0 + reliday | subjID)`). For the hedge version, performance was modeled by a lognormal distribution, while the model for the box version used a Bernoulli distribution with a logit link function. The model computes a correlation between the

participant-specific estimates for each test day. This can be interpreted as the test-retest

reliability. By using this approach, we do not need to compromise on data aggregation and,

therefore, loss of information. Since the model uses hierarchical shrinkage, we obtain

regularized, more accurate person-specific estimates. Most importantly, the model includes

age as a fixed effect. The correlation between the two person-specific estimates is

consequently the age-independent estimate for test-retest reliability. This rules out the

possibility that a high correlation between test days arises from domain general cognitive

development instead of study-specific inter-individual differences. A high correlation

between our participant specific model estimates would speak for a high association

between test days.

Results

We found that our gaze understanding task induced systematic variation: splithalf

and test-retest reliability was high for most samples. For the internal consistency, we show

traditional odd-even splits on our data and the corresponding *Pearson* correlation

coefficients in Figure 2B, C, G and H. Figure 3 compares splithalf reliability coefficients by

splitting and stratification method (Pronk et al., 2021). In the hedge version, the splithalf

reliability coefficients ranged from 0.65 to 0.93. In the box version, splithalf reliability

coefficients ranged from 0.48 to 0.86. Similarly to the results of Pronk et al. (2021), we

found that more robust splitting methods that are less prone to task design or time

confounds yielded higher reliability coefficients. In the majority of cases, stratifying by

target position lead to similar or even higher estimates compared to no stratification. As

might be expected, we found higher coefficients for the samples with higher variation, i.e.,

for our continuous hedge version of the task.

For the test-retest reliability, we show the association between raw performance

scores of the two test days and corresponding *Pearson* correlation coefficients in Figure 2D,
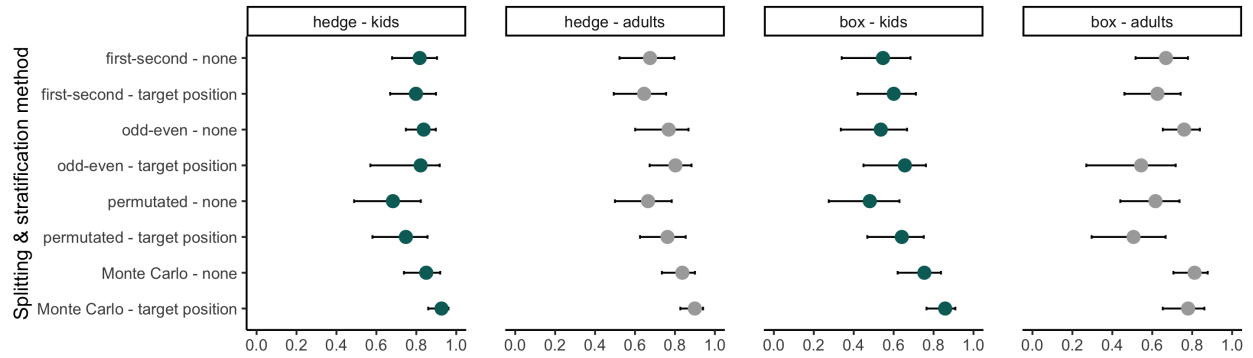
Figure 3. **Internal Consistency**. Reliability coefficients per splitting method, stratification level, study version and age group. Error bars show the 95% confidence intervals of the coefficient estimates, calculated with the function `by_split` from the `splithalfr` package (Pronk et al., 2021).

409   E, I and J.[2]

410   The age-corrected, GLMM based retest reliabilities for children yielded similar results.

411 In the hedge version, the correlation between testdays was 0.90 (95% CI [0.67;1.00]). In the

412 box version, the correlation between testdays was 0.91 (95% CI [0.69;1.00]).

Discussion

414   Our results indicated that the measured variation was systematic. As could be

415 expected, the continuous measure of the hedge version yielded higher reliability estimates

416 than the discrete box version. For children, the model based reliability estimates showed

417 that the task did capture individual differences even when correcting for age. This

418 corroborates what we already see in Figure 2: there was a clear overlap between age

419 groups, indicating that age is predictive of performance for the mean, but is not the main

420 source of individual differences.

---

[2] In the hedge version, we excluded one 3-year-old, one 5-year-old, and two adults from the test-retest analysis. The performance of the mentioned participants was 3 standard deviations above/below the mean of each sample. Including the two children yielded a *Pearson* correlation coefficient of $r = 0.89$. Including the two adults yielded a *Pearson* correlation coefficient of $r = 0.73$.

General Discussion

We have presented a new experimental paradigm to study individual differences in gaze understanding across the lifespan. This paper contributes to methodological advances in developmental psychology in the following ways. First, we could captured fine-grained individual differences in gaze understanding at different ages – from early childhood until adulthood. Individuals behaved consistently different from one another. Second, our task showed satisfactory psychometric properties with respect to internal consistency and retest reliability estimates. Third, our new web-based testing infrastructure proved useful for standardized, portable data collection at scale, both remotely as well as in-person. In sum, our gaze understanding task provides a step towards more robust and reliable research methods, especially with regards to individual differences research.The web app (https://ccp-odc.eva.mpg.de/gafo-demo/) and its source code (https://github.com/ccp-eva/gafo-demo) are freely accessible. We want to highlight that researchers are welcome to use and modify our task according to their needs.

One of the main contributions of the gaze understanding task is that it induces variation between participants. This is primarily achieved through the implementation of short trials, resulting in over a dozen replicates per subject (15 test trials within 5-10 minutes testing time).

In our task, participants are asked to locate a balloon by using the gaze cues of another agent. The outcome measure depends on the presented study version: in the hedge version, we implemented a continuous measure, while the box version employs a categorical measure.

Our continuous measure of children's gaze understanding moves away from treating a social-cognitive ability as an all-or-nothing matter (e.g., dichotomous measures in pass/fail situations) towards an ability on a continuum (Beaudoin et al., 2020; Hughes & Devine, 2015). Identifying variability in social-cognitive abilities is vital for accurately quantifying

developmental change and revealing relations between different aspects of cognition and children's real-life social surroundings. Dedicated measures of individual differences will help us to design meaningful interventions and progress in psychological theory building (Hedge et al., 2018).

Our continuous hedge version yields higher internal consistency estimates compared to the categorical box version. Both study versions exhibit high retest reliability, also when controlling for age. Therefore, when a sufficient amount of trials is presented, the box version of the task can yield reliable individual estimates (cf. Hughes et al. (2000); improved reliability through aggregation). When testing time is limited, we would recommend the use of the continuous study version for higher internal consistency. However, the categorical box version demonstrates design features that might be preferable in some research contexts: researchers could induce different levels of salience for each box. Our task could consequently be used to study bias, preferences and diverse desires (e.g. matching the box appearance to some feature/behavioral characteristic of the agent).

Adding complexity to our gaze understanding task might increase individual differences in an adult sample. Researchers could consider to adjust the task difficulty, for example, by adjusting the target speed or implementing a saccade version (i.e., where the agent only briefly looks at the target position and then back to the center). Another task extension could realize differential feedback dependent on the participant's accuracy. By inducing more between-person variance, reliability estimates could further increase.

As we could see in the splithalf reliability calculations, the more accurately the statistical method represents the task structure, the higher our reliability estimates are. Therefore, we argue that future research should aim at implementing statistical analyses that mirror the experimental design in its complexity. A potentially promising approach are computational cognitive models. We could take advantage of all available information and model variation between and within individuals in an even more fine-grained and

psychologically interpretable manner. Computational frameworks could also be used to model performance and their underlying cognitive processes across tasks. With nested hierarchical models, we could assess the systematic relation between various social-cognitive abilities and recover potentially shared structures between cognitive processes (Bohn, Tessler, Kordt, Hausmann, & Frank, 2022)

In addition to the new task design itself, we designed a new testing infrastructure. Our gaze understanding task is presented as an interactive web app. This enables presentation across devices without any prior installation. Stimuli presentation is achieved through the use of SVGs. This has several advantages: the aspect ratio and stimulus quality are kept constant no matter which size the web browser displays. The cartoon-like presentation makes the task engaging for children and adults alike. Most importantly, we can dynamically modify the stimulus details (e.g., positions) on a trial-by-trial basis. Presented agents and objects can be easily adapted for future task modifications or specific cultural settings.

The web-based implementation allows for different data collection modes: participants can be tested in-person with supervision or remotely at home. Test instructions are standardized and, with prior informed consent, the webcam records study participation. This allows us to scale up data collection: testing is flexible, fast and requires no further experimenter training. We compared children participating in-person, supervised in kindergartens with children participated remotely at home. Our results suggest a comparable developmental trajectory of gaze understanding in both samples. Children in the remote sample are slightly more precise. This effect is most pronounced in the three-year-olds in the box version (for an analysis of the webcam recordings, see Supplements). Therefore, we recommend using a tablet for remote data collection. Children can click for themselves and caregivers have less chance to interfere.

The design choices of the infrastructure underline how our study design can act as a

499 versatile framework for addressing further research questions on social-cognitive

500 development.

501   After having proven satisfactory psychometric properties, the next step will be to

502 ensure the validity of our task. Validity is often assessed by looking at concurrent relations

503 between measures. Studies on gaze following traditionally present children with a

504 simultaneous, congruent movement of eyes and head orientation. In our study, however,

505 children only observe a subtle eye movement. Therefore, our results are not directly

506 comparable to traditional gaze following studies. Another promising way to assess validity

507 is to correlate the social-cognitive ability in question to concepts that are thought to be

508 theoretically related. For example, it has been postulated that social-cognitive abilities are

509 predicted by family-level variables and children's language abilities. In our sample,

510 caregivers were asked to fill out a short questionnaire on children's daily surroundings (e.g.,

511 sibling composition; information on childcare). In addition, a subset of our remote child

512 sample completed a receptive vocabulary task (Bohn, Prein, Delikaya, Haun, & Gagarina,

513 n.d.). We will report the association between these family-level factors and children's

514 language on their gaze understanding abilities in a separate paper (in preparation). In

515 order to match our fine-grained, parametric measure of gaze understanding, however, we

516 hope that future technological advances will enable the continuous measures of children's

517 real-life social surroundings (Barr et al., 2020; Long, Kachergis, Agrawal, & Frank, 2020;

518 Rogoff, Dahl, & Callanan, 2018).

519   Our future goal is to utilize the presented testing infrastructure for further studies on

520 social-cognitive abilities. As has been pointed out, social cognition is encompasses a whole

521 range of abilities which could be best assessed by task batteries (Hughes et al., 2011;

522 Schaafsma et al., 2015). We want to move from the most fundamental social-cognitive

523 abilities to more complex processes like knowledge-ignorance or false belief understanding.

Conclusion

⁵²⁴

We have presented a new experimental paradigm to study individual differences in gaze understanding across the lifespan. The gaze understanding task captures individual differences and shows highly satisfactory psychometric properties with respect to internal consistency and retest reliability. The web-based testing infrastructure allows for standardized, portable data collection at scale, both remotely as well as in-person. Ultimately, this work shows a promising way forward towards more precise measures of cognitive development. The dataset and the analysis code are freely available in the associated online repository (https://github.com/ccp-eva/gazecues-methods). A demo version of the task is available at the following website: https://ccp-odc.eva.mpg.de/gafo-demo/. The implementation architecture (JavaScript and HTML code) and the materials can be accessed in the following repository: https://github.com/ccp-eva/gafo-demo. These resources allow interested researchers to use, extend and adapt the task.

Declarations

Open practices statement

Funding

the children, caregivers and adults who participated in the study. We thank Jana Jurkat for her help with data collection.

Conflicts of interest

The authors declare that they have no conflict of interest.

Consent to participate

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

Authors' contributions

The authors made the following contributions. Julia Prein: Conceptualization, Software, Formal Analysis, Writing - Original Draft Preparation, Writing - Review & Editing; Manuel Bohn: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Steven Kalinke: Software, Writing - Review & Editing; Daniel Haun: Conceptualization, Writing - Review & Editing.

References

Astor, K., Lindskog, M., Forssman, L., Kenward, B., Fransson, M., Skalkidou, A.,
    . . . Gredebäck, G. (2020). Social and emotional contexts predict the
    development of gaze following in early infancy. *Royal Society Open Science*,
    7(9), 201178. https://doi.org/10.1098/rsos.201178

Astor, K., Thiele, M., & Gredebäck, G. (2021). Gaze following emergence relies on
    both perceptual cues and social awareness. *Cognitive Development*, *60*, 101121.
    https://doi.org/10.1016/j.cogdev.2021.101121

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a
    "theory of mind" ? *Cognition*, *21*, 37–46.
    https://doi.org/10.1016/0010-0277(85)90022-8

Barr, R., Kirkorian, H., Radesky, J., Coyne, S., Nichols, D., Blanchfield, O., . . .
    Fitzpatrick, caroline. (2020). Beyond Screen Time: A Synergistic Approach to a
    More Comprehensive Assessment of Family Media Exposure During Early
    Childhood. *Frontiers in Psychology*, *11*.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., . . .
    Laurence, S. (2013). Early false-belief understanding in traditional non-Western
    societies. *Proceedings of the Royal Society B: Biological Sciences*.
    https://doi.org/10.1098/rspb.2012.2654

Beaudoin, C., Leblanc, É., Gagner, C., & Beauchamp, M. H. (2020). Systematic
    Review and Inventory of Theory of Mind Measures for Young Children.
    Frontiers in Psychology, *0*. https://doi.org/10.3389/fpsyg.2019.02905

Benson, J., Sabbagh, M., Carlson, S., & Zelazo, P. (2012). Individual Differences in
    Executive Functioning Predict Preschoolers' Improvement From Theory-of-Mind
    Training. *Developmental Psychology*, *49*. https://doi.org/10.1037/a0031056

Bohn, M., Prein, J., Delikaya, B., Haun, D. B. M., & Gagarina, N. (n.d.). *oREV:*
    An Item Response Theory based open receptive vocabulary task for 3 to

588   *8-year-old children.*

589   Bohn, M., Tessler, M. H., Kordt, C., Hausmann, T., & Frank, M. C. (2022). *An*

590   *individual differences perspective on the development of pragmatic abilities in the*

591   *preschool years.* PsyArXiv. https://doi.org/10.31234/osf.io/s2e3p

592   Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its

593   relation to language. *Developmental Science*, *8*(6), 535–543.

594   https://doi.org/10.1111/j.1467-7687.2005.00445.x

595   Bulgarelli, D., & Molina, P. (2016). Social Cognition in Preschoolers: Effects of

596   Early Experience and Individual Differences. *Frontiers in Psychology*, *7*.

597   https://doi.org/10.3389/fpsyg.2016.01762

598   Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using

599   Stan. *Journal of Statistical Software*, *80*(1).

600   https://doi.org/10.18637/jss.v080.i01

601   Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package

602   brms. *The R Journal*, *10*(1), 395. https://doi.org/10.32614/RJ-2018-017

603   Buttelmann, D., Kühn, K., & Zmyj, N. (2021). The Relations among Theory of

604   Mind, Inhibitory Control, and Aggressive Behavior in 4-Year-Old Children – A

605   Multi-Measure Multi-Informant Approach. *Journal of Cognition and*

606   *Development*, *0*(0), 1–24. https://doi.org/10.1080/15248372.2021.1987240

607   Byers-Heinlein, K., Tsui, R. K.-Y., van Renswoude, D., Black, A. K., Barr, R.,

608   Brown, A., . . . Singh, L. (2021). The development of gaze following in

609   monolingual and bilingual infants: A multi-laboratory study. *Infancy*, *26*(1),

610   4–38. https://doi.org/10.1111/infa.12360

611   Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., . . .

612   Singh, S. (2005). Synchrony in the onset of mental-state reasoning: Evidence

613   from five cultures. *Psychological Science.*

614   https://doi.org/10.1111/j.0956-7976.2005.01544.x

Carlson, S., & Moses, L. J. (2001). Individual Differences in Inhibitory Control and
        Children's Theory of Mind. *Child Development*, *72*(4), 1032–1053.

Carlson, S., Moses, L. J., & Claxton, L. J. (2004). Individual differences in
        executive functioning and theory of mind: An investigation of inhibitory control
        and planning ability. *Journal of Experimental Child Psychology*, *87*(4), 299–319.
        https://doi.org/10.1016/j.jecp.2004.01.002

Coelho, E., George, N., Conty, L., Hugueville, L., & Tijus, C. (2006). Searching for
        asymmetries in the detection of gaze contact versus averted gaze under different
        head views: A behavioural study. *Spatial Vision*, *19*(6), 529–545.
        https://doi.org/10.1163/156856806779194026

Cutting, A. L., & Dunn, J. (1999). Theory of Mind, Emotion Understanding,
        Language, and Family Background: Individual Differences and Interrelations.
        Child Development, *70*(4), 853–865. https://doi.org/10.1111/1467-8624.00061

Del Bianco, T., Falck-Ytter, T., Thorup, E., & Gredebäck, G. (2019). The
        Developmental Origins of Gaze-Following in Human Infants. *Infancy*, *24*(3),
        433–454. https://doi.org/10.1111/infa.12276

Dunn, J., Brown, J., Slomkowski, C., Tesla, C., & Youngblade, L. (1991). Young
        children's understanding of other people's feelings and beliefs: Individual
        differences and their antecedents. *Child Development*, *62*(6), 1352–1366.

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention:
        Visual attention, social cognition, and individual differences. *Psychological
        Bulletin*, *133*(4), 694–724. https://doi.org/10.1037/0033-2909.133.4.694

Gola, A. A. H. (2012). Mental verb input for promoting children's theory of mind:
        A training study. *Cognitive Development*, *27*(1), 64–76.
        https://doi.org/10.1016/j.cogdev.2011.10.003

Gopnik, A., & Slaughter, V. (1991). Young Children's Understanding of Changes in
        Their Mental States. *Child Development*, *62*(1), 98.

https://doi.org/10.2307/1130707

Happé, F., Cook, J. L., & Bird, G. (2017). The Structure of Social Cognition: In(ter)dependence of Sociocognitive Processes. *Annual Review of Psychology*, *68*(1), 243–267. https://doi.org/10.1146/annurev-psych-010416-044046

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hernik, M., & Broesch, T. (2019). Infant gaze following depends on communicative signals: An eye-tracking study of 5- to 7-month-olds in Vanuatu. *Developmental Science*, *22*(4), e12779. https://doi.org/10.1111/desc.12779

Herrmann, E., Hernández-Lloreda, M. V., Call, J., Hare, B., & Tomasello, M. (2010). The Structure of Individual Differences in the Cognitive Abilities of Children and Chimpanzees. *Psychological Science*, *21*(1), 102–110. https://doi.org/10.1177/0956797609356511

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good Test-Retest Reliability for Standard and Advanced False-Belief Tasks across a Wide Range of Abilities. *Journal of Child Psychology and Psychiatry*, *41*(4), 483–490. https://doi.org/10.1111/1469-7610.00633

Hughes, C., & Devine, R. T. (2015). Individual Differences in Theory of Mind From Preschool to Adolescence: Achievements and Directions. *Child Development Perspectives*, *9*(3), 149–153. https://doi.org/10.1111/cdep.12124

Hughes, C., & Ensor, R. (2007). Executive function and theory of mind: Predictive relations from ages 2 to 4. *Developmental Psychology*, *43*(6), 1447–1459. https://doi.org/10.1037/0012-1649.43.6.1447

Hughes, C., Ensor, R., & Marks, A. (2011). Individual differences in false belief understanding are stable from 3 to 6 years of age and predict children's mental state talk with school friends. *Journal of Experimental Child Psychology*, *108*(1),

96–112. https://doi.org/10.1016/j.jecp.2010.07.012

Hughes, C., & Leekam, S. (2004). What are the Links Between Theory of Mind and Social Relations? Review, Reflections and New Directions for Studies of Typical and Atypical Development. *Social Development*, *13*(4), 590–619. https://doi.org/10.1111/j.1467-9507.2004.00285.x

Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., & Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental Psychology*, *52*(8), 1192–1205. https://doi.org/10.1037/dev0000140

Itakura, S., & Tanaka, M. (1998). Use of experimenter-given cues during object-choice tasks by chimpanzees (Pan troglodytes), an orangutan (Pongo pygmaeus), and human infants (Homo sapiens). *Journal of Comparative Psychology*, *112*(2), 119–126. https://doi.org/10.1037/0735-7036.112.2.119

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences*, *22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Lecce, S., Bianco, F., Devine, R. T., Hughes, C., & Banerjee, R. (2014). Promoting theory of mind during middle childhood: A training program. *Journal of Experimental Child Psychology*, *126*, 52–67. https://doi.org/10.1016/j.jecp.2014.03.002

Lee, K., Eskritt, M., Symons, L. A., & Muir, D. (1998). Children's use of triadic eye gaze information for "mind reading". *Developmental Psychology*, *34*(3), 525–539. https://doi.org/10.1037//0012-1649.34.3.525

Long, B., Kachergis, G., Agrawal, K., & Frank, M. C. (2020). *Detecting social information in a dense dataset of infants' natural visual experience* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/z7tdg

Mayer, A., & Träuble, B. (2015). The weird world of cross-cultural false-belief

research: A true- and false-belief study among samoan children based on commands. *Journal of Cognition and Development*, *16*(4), 650–665. https://doi.org/10.1080/15248372.2014.926273

Mayes, L. C., Klin, A., Tercyak, K. P., Cicchetti, D. V., & Cohen, D. J. (1996). Test-Retest Reliability for False-Belief Tasks. *Journal of Child Psychology and Psychiatry*, *37*(3), 313–319. https://doi.org/10.1111/j.1469-7610.1996.tb01408.x

McEwen, F., Happé, F., Bolton, P., Rijsdijk, F., Ronald, A., Dworzynski, K., & Plomin, R. (2007). Origins of individual differences in imitation: Links with language, pretend play, and socially insightful behavior in two-year-old twins. *Child Development*, *78*(2), 474–492. https://doi.org/10.1111/j.1467-8624.2007.01010.x

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-belief Understanding. *Child Development*, *78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Molleman, L., Kurvers, R. H. J. M., & van den Bos, W. (2019). Unleashing the BEAST: A brief measure of human social information use. *Evolution and Human Behavior*, *40*(5), 492–499. https://doi.org/10.1016/j.evolhumbehav.2019.06.005

Moore, C. (2008). The Development of Gaze Following. *Child Development Perspectives*, *2*(2), 66–70. https://doi.org/10.1111/j.1750-8606.2008.00052.x

Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., & Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child Development*, *78*(3), 938–954. https://doi.org/10.1111/j.1467-8624.2007.01042.x

Okumura, Y., Kanakogi, Y., Kobayashi, T., & Itakura, S. (2017). Individual differences in object-processing explain the relationship between early gaze-following and later language development. *Cognition*, *166*, 418–424.

https://doi.org/10.1016/j.cognition.2017.06.005

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Pavarini, G., de Hollanda Souza, D., & Hawk, C. K. (2013). Parental Practices and Theory of Mind Development. *Journal of Child and Family Studies*, *22*(6), 844–853. https://doi.org/10.1007/s10826-012-9643-8

Peterson, Wellman, H. M., & Slaughter, V. (2012). The Mind Behind the Message: Advancing Theory-of-Mind Scales for Typically Developing Children, and Those With Deafness, Autism, or Asperger Syndrome: **The Mind Behind the Message**. *Child Development*, *83*(2), 469–485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2021). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-01948-3

R Core Team. (2022). *R: A language and environment for statistical computing* [Manual]. Vienna, Austria: R Foundation for Statistical Computing.

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, 1–13. https://doi.org/10.1038/s44159-022-00037-z

Repacholi, V. S. &. B. (2003). Introduction Individual Differences in Theory of Mind: What Are We Investigating? In *Individual Differences in Theory of Mind*. Psychology Press.

Rizzo, M. T., & Killen, M. (2018). Theory of mind is related to children's resource allocations in gender stereotypic contexts. *Developmental Psychology*, *54*(3), 510. https://doi.org/10.1037/dev0000439

Rogoff, B., Dahl, A., & Callanan, M. (2018). The importance of understanding children's lived experience. *Developmental Review*, *50*, 5–15. https://doi.org/10.1016/j.dr.2018.05.006

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*(2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007

Shepherd, S. (2010). Following Gaze: Gaze-Following Behavior as a Window into Social Cognition. *Frontiers in Integrative Neuroscience*, *4*.

Slaughter, V. (2015). Theory of Mind in Infants and Young Children: A Review. *Australian Psychologist*, *50*(3), 169–172. https://doi.org/10.1111/ap.12080

Sodian, B., Licata, M., Kristen-Antonow, S., Paulus, M., Killen, M., & Woodward, A. (2016). Understanding of Goals, Beliefs, and Desires Predicts Morally Relevant Theory of Mind: A Longitudinal Investigation. *Child Development*, *87*(4), 1221–1232. https://doi.org/10.1111/cdev.12533

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis. *Journal of Human Evolution*, *52*(3), 314–320. https://doi.org/10.1016/j.jhevol.2006.10.001

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist*, *30*(2), 128–134. https://doi.org/10.1037/h0076759

Walker, S. (2005). Gender Differences in the Relationship Between Young Children's Peer-Related Social Competence and Individual Differences in Theory of Mind. *The Journal of Genetic Psychology*, *166*(3), 297–312. https://doi.org/10.3200/GNTP.166.3.297-312

Wellman, H. M. (2012). Theory of mind: Better methods, clearer findings, more development. *European Journal of Developmental Psychology*, *9*(3), 313–330. https://doi.org/10.1080/17405629.2012.680297

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*, 655–684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind Tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*. https://doi.org/10.1016/0010-0277(83)90004-5
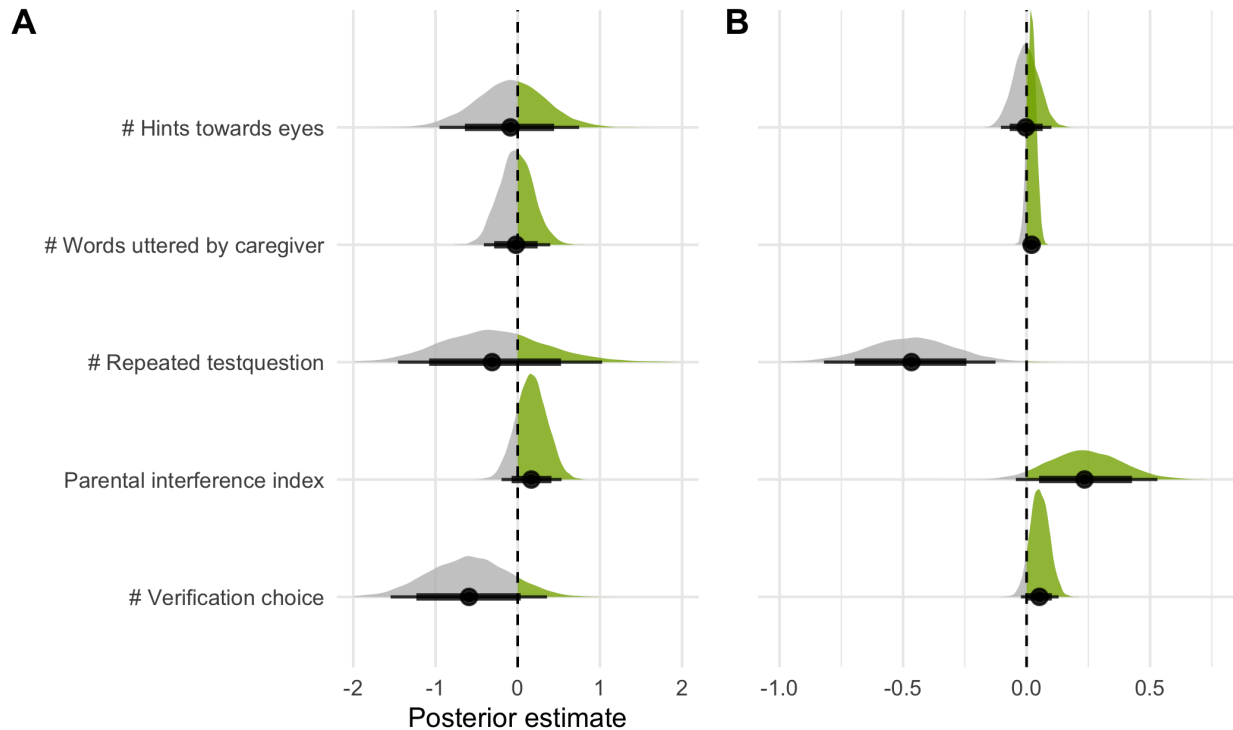
Supplements

Child sample

Figure 4. **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).

790   **Webcam coding.**   Comparing the performances of children across our two data

791 collection modes, we found that children participating remotely were slightly more precise.

792 This difference was especially prominent in younger participants in the box version of the

793 task. It is conceivable that caregivers were especially prone to influence the behavior of

794 younger children. In the box version, caregivers might have had more opportunities to

795 interfere since they carried out the clicking for their children. In an exploratory analysis,

796 we coded parental behavior and environmental factors during remote unsupervised testing.

797 Due to the time consuming nature of hand coding videos frame by frame, we focused on

the subsample with the greatest performance difference between data collection modes: the three-year-olds in the box version of the task (n = 16). We reasoned that if parental 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 + (1 | subjID) + (1 | targetPosition)`.[3]

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

---

[3] Attentive readers might notice that we simplified the structure of random effects. Compared to our models in the *Individual differences* section, this model does not include the random slope for symmetric target position within participants. We decided to do so since we had limited amount of data from few participants.

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

caregivers asked for children's responses again, the more likely children clicked on the

correct box. Interestingly, the effect reversed on a subject level - possibly due to greater

learning effects for the children that were most likely to click incorrectly in the beginning

and then receiving most parental comments. On the subject level, the model including

number of repeated test questions performed best: the more caregivers asked again where

the target landed, the more likely children were to respond to the incorrect box. In all

cases, however, ELPD difference scores were smaller than their standard errors. Similarly,

95% CI of the model estimates included zero and were rather wide (Table 1). Therefore, we

conclude that the effect of parental interference was negligable and could, most likely, be

explained as described above.

Adult sample

Recruitment.   We recruited participants using the online participant recruitment

service *Prolific* from the University of Oxford. *Prolific*'s subject pool consists of a mostly

European and US-american sample although subjects from all over the world are included.

The recruitment platform realises ethical payment of participants, which requires

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

estimated average time taken to complete the task. *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. Therefore, study data

Table 1

Model comparison for exploratory webcam coding of parental interference

Predictor	WAIC	Weight	ELPD	SE
By trial - # Verification choice	262.98	0.57	0.00	0.00
By trial - Null model	263.25	0.43	-0.13	1.19
By trial - Parental interference index	264.32	0.00	-0.67	0.97
By trial - # Repeated testquestion	265.34	0.00	-1.18	1.26
By trial - # Words uttered by caregiver	266.27	0.00	-1.65	1.21
By trial - # Hints eyes	266.33	0.00	-1.68	1.27
By subject - # Repeated testquestion	83.83	0.70	0.00	0.00
By subject - Parental interference index	88.50	0.00	-2.34	3.56
By subject - Null model	89.62	0.00	-2.90	3.60
By subject - # Verification choice	89.75	0.30	-2.96	4.46
By subject - # Words uttered by caregiver	90.46	0.00	-3.32	4.21
By subject - # Hints eyes	92.90	0.00	-4.54	3.73

Note. All models included random intercepts for participant and target
position.

was saved only on our internal servers, while *Prolific* provided demographic information of

the participants. Participants' *Prolific* ID was forwarded to our study website using URL

parameters. This way, we could match participant demographic data to our study data.

The same technique was used to confirm study completion: we redirected participants from

our study website back to the *Prolific* website using URL parameters. We used *Prolific*'s

inbuilt prescreening filter to include only participants who were fluent in English and could

therefore properly understand our written and oral study instructions.

Study 1 - Validation hedge version.   The aim of Study 1 was to validate the

hedge version of our gaze understanding 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 *Prolific*, and employed the same methodology. However, this time we focussed on 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 withdrew their submission after study completion. Data collection took place in June 2021. Participants were compensated with £1.00 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 04:43min to complete the study.

Study 3 - Reliability hedge version.   In study 3 and 4, we assessed the test-retest reliability of our gaze understanding 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

888   participate in a study. This way, repeated measurements can be implemented, collecting

889   data from the same subjects at different points in time.

890   In a first round, 60 participants took part on the first testday. Additional two

891   subjects returned their submission, i.e., decided to leave the study early or withdrew their

892   submission after study completion. One additional participant timed out, i.e., did not

893   finish the survey within the allowed maximum time. The maximum time is calculated by

894   *Prolific*, based on the estimated average completion time. For this study, the maximum

895   time amounted to 41 minutes. For the first testday, participants were compensated with

896   £1.25. We estimated an average completion time of 9 minutes, resulting in an estimated

897   hourly rate of £8.33. In average, participants took 07:11min to complete the first part.

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

899   additional participant timed out, i.e., did not finish the survey within the allowed

900   maximum time. Participants were compensated with £1.50 for completing the second part

901   of the study. We estimated an average completion time of 9 minutes, resulting in an

902   estimated hourly rate of £10. In average, participants took 06:36min to complete the

903   second part of the study.

904   Since we aimed for a minimum sample size of 60 subjects participating on both

905   testdays, we reran the first testday with additional 50 participants. Additional seven

906   subjects returned their submission, i.e., decided to leave the study early or withdrew their

907   submission after study completion. Two additional participants timed out, i.e., did not

908   finish the survey within the allowed maximum time. Again, participants were compensated

909   with £1.25 for completing the first part of the study (estimated average completion time 9

910   minutes, estimated hourly rate of £8.33). In average, participants took 06:51min to

911   complete the first part.

912   Of the additional 50 participants that completed testday 1, 29 subjects finished

913   testday 2. Again, participants were compensated with £1.50 for completing the second

914 part of the study (estimated average completion time 9 minutes, estimated hourly rate of

915 £10). In average, participants took 06:26min to complete the second part of the study.

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

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

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

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

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

921 In a first round, 60 participants took part on the first testday. Additional five

922 subjects returned their submission, i.e., decided to leave the study early or withdrew their

923 submission after study completion. For the first testday, participants were compensated

924 with £1.25. We estimated an average completion time of 9 minutes, resulting in an

925 estimated hourly rate of £8.33. In average, participants took 07:33min to complete the

926 first part.

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

928 Participants were compensated with £1.50 for completing the second part of the study. We

929 estimated an average completion time of 9 minutes, resulting in an estimated hourly rate of

930 £10. In average, participants took 07:50min to complete the second part of the study.

931 Since we aimed for a minimum sample size of 60 subjects participating on both

932 testdays, we reran the first testday with additional 50 participants. Additional eight

933 subjects returned their submission, i.e., decided to leave the study early or withdrew their

934 submission after study completion. One additional participant timed out, i.e., did not

935 finish the survey within the allowed maximum time. Again, participants were compensated

936 with £1.25 for completing the first part of the study (estimated average completion time 9

937 minutes, estimated hourly rate of £8.33). In average, participants took 07:37min to

938 complete the first part.

939 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.

Instructions and voice over descriptions

This is the content of our audio recordings that were played as instructions and 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! The animal helps you and looks at the balloon.	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