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

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

[1] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Author Note

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

Abstract

In order to explain and predict the behavior of agents, we use social cognition: we represent and reason about other's perspectives, knowledge, intentions, beliefs and preferences. Traditional measures of social cognition (e.g., false belief change-of-location tasks), however, 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 on 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. To approach this issue, we designed a balloon finding task to study social cognition in young children and adults. We concentrate on an essential ability that is involved in many social-cognitive reasoning processes: gaze cue understanding – the ability to locate and use the attentional focus of an agent. Our interactive web interface works across devices and enables supervised and unsupervised, as well as in-person and remote testing. The implemented spatial layout allows for discrete as well as continuous measures of participants' click imprecision and is easily adoptable to different study needs. Here we show that our task induces inter-individual differences in a child (N = XXX) and 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-test reliability estimates underline that the capured variation is systematic. Furthermore, we find first evidence for the external validity of our task: the measured performance in gaze cue understanding relates to children's real life social surrounding. Taken together, 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, psychometrics

Word count: X

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

Maybe for abstract: 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 individual differences.

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 best-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.

But we still know relatively little about the co-development of social-cognitive abilities, variation in the individual's pace of mastering developmental milestones, and which factors influence a child's proficiency.

Researchers are interested in questions concerning individual differences in the social cognition domain (Hughes & Devine, 2015). Previous studies focused, for example, on the association of stereotypes and theory of mind (Rizzo & Killen, 2018), mental state talk and false belief understanding (Hughes, Ensor, & Marks, 2011), attachment quality and gaze following (Astor et al., 2020), language and gaze following (Okumura, Kanakogi, Kobayashi, & Itakura, 2017), emotion, language, family interaction and belief understanding (Bulgarelli & Molina, 2016; Cutting & Dunn, 1999; Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991), and the interplay of physical and social cognition (Herrmann, Hernández-Lloreda, Call, Hare, & Tomasello, 2010). Correlational approaches have also been used to argue for and test assumptions of cognitive development theories (Kidd, Donnelly, & Christiansen, 2018; Mundy et al., 2007 May-Jun; Underwood, 1975).

This is due to the fact that many developmental psychology studies track change at an (age-) group level. Traditional measures of social cognition are not designed to capture variation between children: they often rely on low trial numbers, small sample sizes, dichotomous measures, and consequently lack satisfactory psychometric properties. 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 must take individual differences seriously.

We need tasks that capture variability in performance reliably (i.e., systematically order individuals in the same way) and meaningfully (i.e., differences in test scores correspond to differences in ability).

A recent paper by Hedge, Powell, and Sumner (2018) argued that cognitive tasks often become well-established by displaying robust effects across studies. The authors could show that this often does not emerge due to high measurement variance but rather low between-subject variability, concluding that popular cognitive tasks may reliably measure group differences but not individual differences. In a similar vein, Pronk, Molenaar, Wiers, and Murre (2021) reasoned that accurate reliability estimates are needed to judge how well a certain cognitive test is suited to draw inferences about individuals.

[TODO: check Manuels references]

However, we rarely find studies addressing these issues. Methods, that have currently been used, are… One of the most commonly applied measures in social cognition research is the Sally-Anne task. [TODO: fill in FB studies] The outcome measures false belief understanding in a dichotomous way: children pass the task if they answer that the protagonist would … With four to five years of age, most children pass the task. From this age onwards, the change-of-location task shows ceiling effects and has very limited diagnostic value. With these procedures, researchers could systematically determine the average age at which children succeed in the Sally-Anne task. However, no individual differences can be visible, especially when applying only one trial per child.

Our aim is, therefore, to make individual differences measurable in a systematic and reliable way. For getting precise person-specific estimates, we need an accurate measure of a child's given underlying social-cognitive ability. One of our main goals is to create a task that induces variation between children. To approach this issue, we focus on a continuous measure and short trials that facilitate more than a dozen replicates per subject.

To estimate the effect of inter-individual differences, we need a large sample size since some of the effects that we are looking for might be small. Since data collection with families requires a lot of organizational effort, we want to enable data collection in-person with supervision as well as remotely without supervision. A standardized, easily accessible test procedure helps us to enable data collection at scale.

Another goal in creating new tasks should be to focus on the face value: we want to measure the underlying social-cognitive ability as straight-forward and directly as possible. Importantly, we can assess whether the task shows external validity by assessing the relationship with related constructs and factors of children's social surroundings.

In our first novel task, we concentrated on the fundamental ability that is involved in many social-cognitive reasoning processes: gaze cue understanding - the ability to locate and use the attentional focus of an agent. The first component is often termed gaze following - turning ones eyes in the same direction as the gaze of another agent - and has been studied intensively [TODO: add references]. By acting as a 'front end ability' (Brooks & Meltzoff, 2005, p. 535), following an agent's gaze provides insights into their intentions, thoughts, and feelings. In our definition, gaze cue understanding goes one step further by including the *acting* on the gaze-cued location.

- why is social cognition important
- what methods are currently been used?: wellman
- what are common issues?
- what to aim at. individual differences in developmental psychology
- what characteristics should a new task fulfill? reliable tasks, variation needed, more trials
- goal of the current project: standardized, easy to use continuous methods
- soc cog, gaze cue understanding small def (longer one dev paper)

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:

“ToM also has specific real-life consequences. First, the development of ToM competence goes along with general measures of children’s peer social skills in early and middle childhood. (. . .) Second, ToM specifically predicts communicative competence. (. . .) Third, ToM competence is related to the quality of peer relationships: children with more advanced ToM are rated as more likeable and popular among their peers. Fourth, children who are more proficient at ToM tasks tend to act more prosocially, including comforting, sharing or helping other individuals. Finally, preschool ToM competence predicts achievement in primary school, a relationship that is possibly mediated by social competence, in that preschool ToM abilities enable subsequent social competence development, which in turn contributes to school achievement.” (p. 2) “Evidence for an emerging understanding of perception at 9 months of age comes from various sources. For example, children begin to follow the gaze of other agents in systematic and differential ways: they follow an agent’s head turn only when the agent can actually see (has their eyes open rather than closed, or wears a transparent rather than an opaque blindfold).” (p. 2) Developmental determinants: executive function, language (“that” complementations), social (SES, siblings, mind-minded parents) Implicit tasks: “A third class of implicit ToM tasks is interaction tasks, in which participants are involved in a communicative or cooperative interaction with another agent. This agent forms a mental state (such as a true or false belief regarding the contents of a box) and experimenters measure whether participants spontaneously take the agent’s belief into account in their interaction with the agent (for instance by helping or by interpreting the agent’s communicative acts accordingly)” (p. 9) => reliable & valid tasks to assess coherent development of social-cognitive functions

Task design

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. An 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`, `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 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: they respond by touching or clicking on the screen. Visual access to the target's true location is manipulated by a hedge. Participants either have full, partial, or no visual access to the true target location. When partial or no information about the target location is accessible, participants are expected to use the agent's gaze as a cue.

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?". For confirming the participant's click, a short plop sound plays and a small orange circle appears at the location of choice. If no response is registered within 5 secs after the target landed, an audio prompt reminds the participant to respond.

Trials

Trials differ in the amount of visual access that participants have to the final target position. Before the test trials start, participants complete four training trials during which they familiarize themselves with clicking the screen. In the first training trial, participants have full visual access to the target flight and the target's end location and are simply asked to click on the visible balloon. In the second and third training trials, participants have partial access: they witness the target flight but cannot see the target's end location. They are then asked to click on the hidden balloon, i.e., the location where they saw the target land. In test trials, participants have no visual access to the target flight or the end location. Participants are expected to use the agent's gaze as a cue to locate the target. The first trial of each type comprises a voice-over description of the presented events. The audio descriptions explicitly state that the agent is always looking at the target (see Appendix for audio script). After the four training trials, participants receive 15 test trials. The complete sequence of four training trials and 15 test trials can be easily completed within 5-10 minutes.

Study versions

We designed two study versions which differ in the final hiding place of the target and, consequently, on the outcome measure: a *hedge version* (continuous) and a *box version* (discrete). 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. Here, the dependent variable is imprecision, which is defined as the absolute difference between the target center and the x coordinate of the participant's click. In the box version, the target lands in a box and participants are asked to click on the box that hides the target. Researchers have the choice 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 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 (see Figure 1B and C).

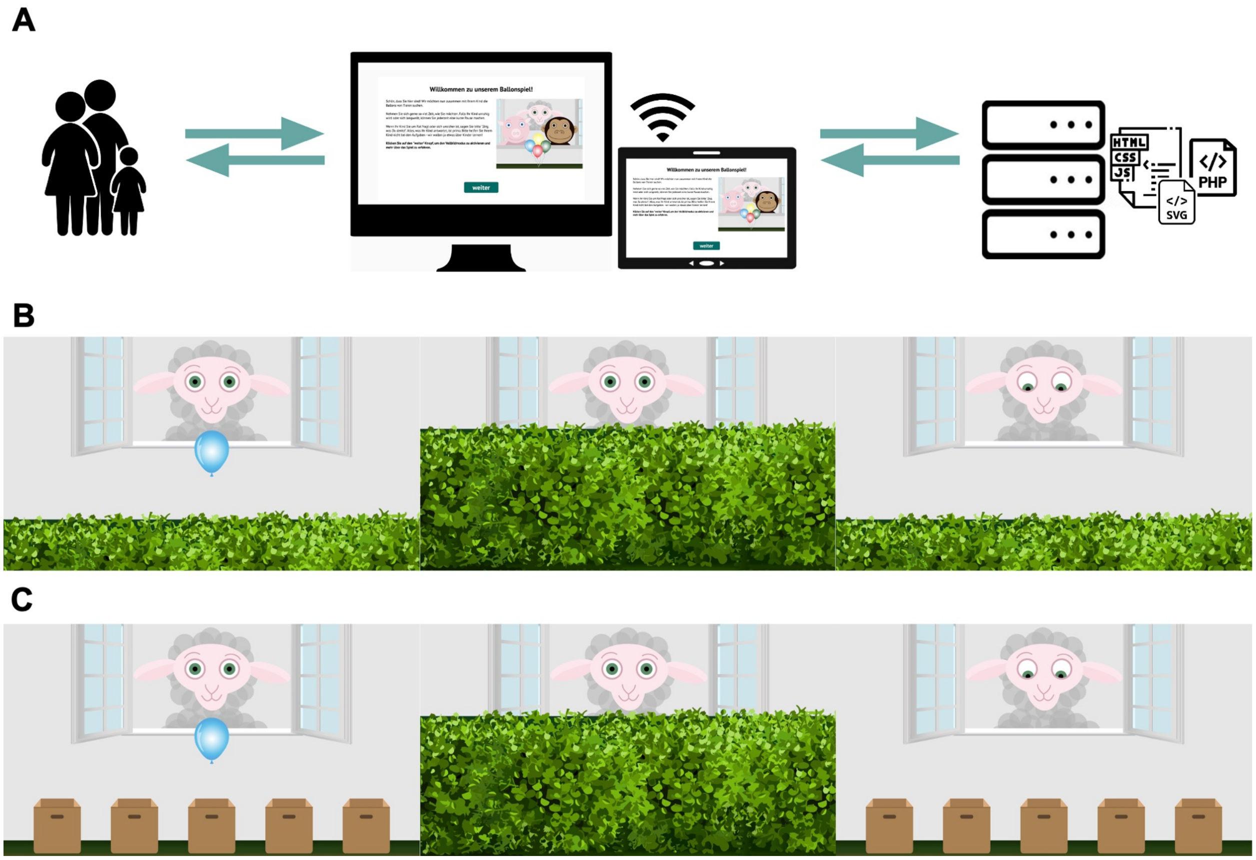

Figure 1. **Study setup**. (A) Infrastructure for online testing. (i) Subjects aged 3 – 99+ can participate. Data collection can take place anywhere: online, in kindergartens or research labs. (ii) The task is presented as a website that works across devices. (iii) The scripts for the website and the recorded data are stored on secure local servers. (B) Hedge version (continuous) of the balloon finding 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 balloon finding task. Number of boxes (min. 1; max. 8) as potential hiding locations can be set according to the researcher's need.

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.

Individual differences

Our first aim was to assess whether our balloon finding task induces inter-individual variation in a child and adult sample. Furthermore, we were interested in how the data collection mode 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. 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*. *Prolific* is an online participant recruitment service from the University of Oxford with a predominantly European and US-american subject pool. Participants consisted of 50 and 50 English-speakers with an average age of 31.92 and 30.76 years (SD = 12.15 and 9.12, range = 18 and 19 - 63 and 59, 36 and 28 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 cue understanding and the effect of data collection mode, we fit a GLMM predicting the task performance 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 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

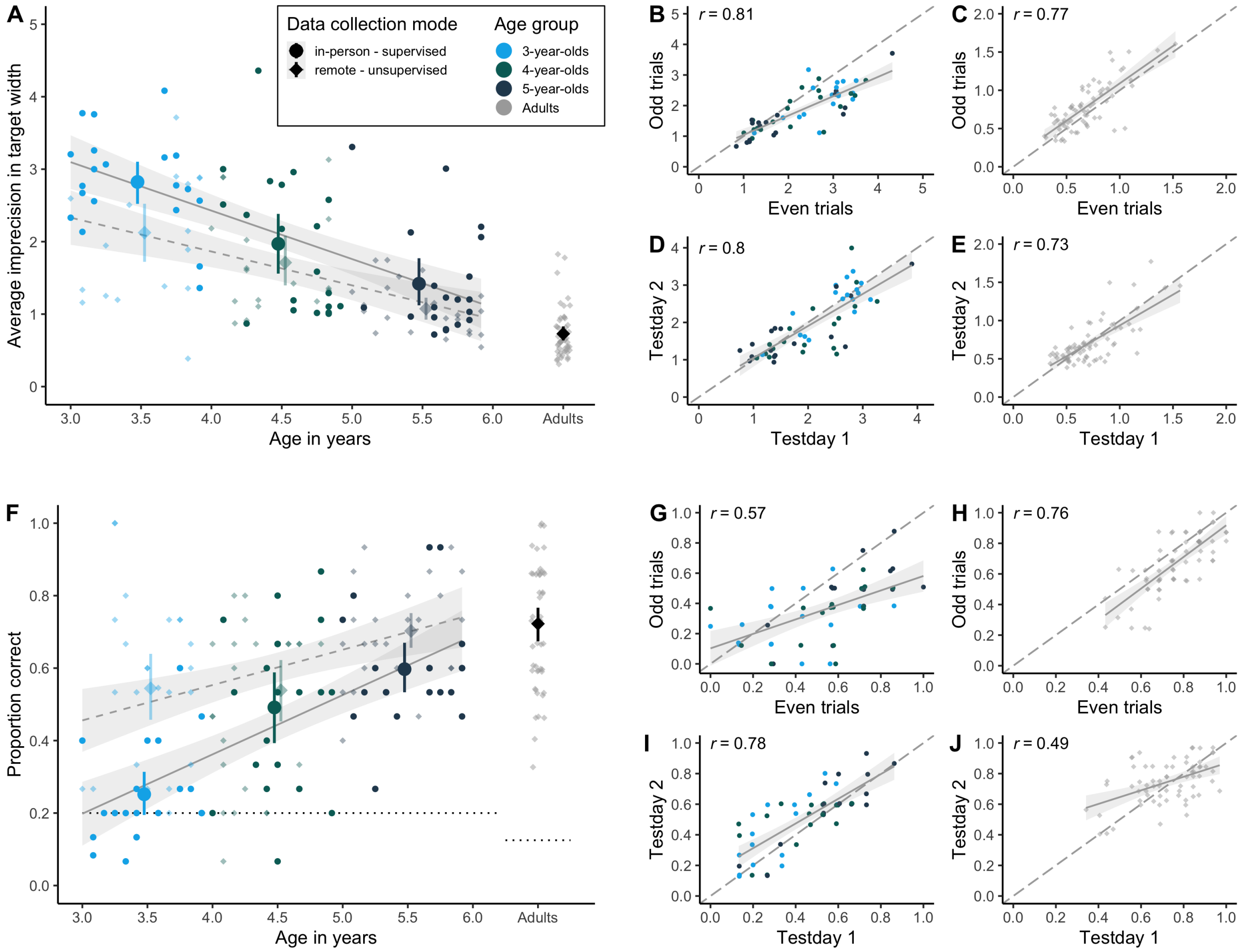

Figure 2. **Measuring inter-individual variation**. (A) Developmental trajectory in continuous hedge version. Performance is measured as average imprecision, i.e., the absolute distance between the target's center and the participant's click. The unit of imprecision is counted in the width of the target, i.e., a participant with 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. 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. 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.

the clicking for their children.[1]

Our GLMM analysis corroborated the visual inspection of the data: in the hedge version, the estimates for age (β = -0.32; 95% CI [-0.41; -0.24]) and data collection mode -0.32 (95% CI [-0.49; -0.14]) were negative and reliably different from zero. In the box version, the estimate of age (β =0.63 (95% CI [0.40; 0.88]) and the estimate of data collection mode (β = 1.12 (95% CI [0.69; 1.57]) were positive and reliably different from zero. Note that even though confidence intervals from the data collection estimates were wide, the effect was positive and reliably different from zero in a way that our remote sample performed more accurately than our in-person sample.

Discussion

Our task induced inter-individual variation in both adults and children. We see substantial developmental gains: with increasing age, participants got more and more precise in locating the target. The five-year-olds reached a proficiency level close to the adults' level. For neither study version nor age group did we find any floor or ceiling effects. The presentation as a tablet game kept children interest and motivated throughout the 15 test trials. Furthermore, we found a comparable developmental trajectory for an unsupervised remote child sample. This illustrates the flexibility of the task.

Internal consistency and retest reliability

As a next step, we aimed at investigating whether the variation that we captured with our balloon finding task is reliable. We assessed internal consistency (split-half reliability) and test-retest reliability. Data collection and analysis were pre-registered (can

[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 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.

be found here: https://osf.io/xqm73 for child sample, and https://osf.io/nu62m 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 July 2021 and April 2022.

Participants

Participants were recruited in the same way as in the previous study. The child sample consisted of 106 children, including 35 3-year-olds (mean = 42.57 months, SD = 2.98, range = 38 - 47, 17 girls), 38 4-year-olds (mean = 53.77 months, SD = 3.16, range = 48 - 59, 20 girls), and 33 5-year-olds (mean = 66.12 months, SD = 3.36, range = 61 - 71, 17 girls).

The adult sample consisted of 70 and 66 English-speakers with an average age of 25.43 and 26.05 years (SD = 6.43 and 9.44, range = 18 and 18 - 51 and 71, 45 and 42 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 et al. (2021) recently compared various methods for computing split-half 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 balloon finding 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 0.57 to 0.84. In the box version, splithalf reliability coefficients ranged from 0.49 to 0.76. 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,

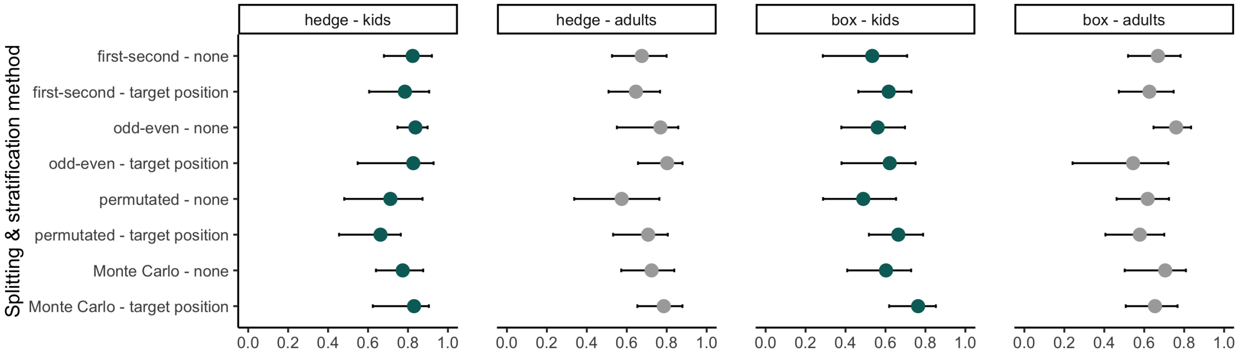

Figure 3. **Internal Consistency**. Reliability coefficients per splitting method, stratification level, study version and age group. Error bars show the 95% confidence intervals of the coefficient estimates, calculated with the function `by_split` from the `splithalfr` package (Pronk et al., 2021).

E, I and J.[2]

The age-corrected, GLMM based retest reliabilities for children yielded similar results. In hedge version it was 0.90 (95% CI [0.68;1.00]). In the box version it was 0.92 (95% CI [0.70;1.00]).

Discussion

Our results indicated that the measured variation was systematic. As could be expected, the continuous measure of the hedge version yielded higher reliability estimates than the discrete box version. For children, the model based reliability estimates showed that the task did capture individual differences even when correcting for age. This corroborates what we already see in Figure 2: there was a clear overlap between age groups, indicating that age is predictive of performance for the mean, but is not the main source of individual differences.

[2] In the hedge version, we excluded one 5-year-old child from the test-retest analysis. The performance of the mentioned child was 3 standard deviations above the mean on both test days. Including the child yielded a *Pearson* correlation coefficient of $r = 0.87$.

Validity

Our third aim was to assess whether the captured individual variation in gaze cue understanding relates to factors in children's real live social environment. Previous studies found associations between social cognition measures and various environmental factors (Devine & Hughes, 2018), including family background and education [Cutting and Dunn (1999); bulgarelli2016social], number and age of siblings and family constellation Cassidy, Fineberg, Brown, & Perkins (2005), interaction with siblings (Dunn et al., 1991), and centre-based childcare (Bulgarelli & Molina, 2016). It is assumed that opportunities to play, communicate and argue with siblings and similarly-aged peers help children to understand the human mind. Therefore, if we find a link between gaze cue understanding and family factors, we regard this as an indicator for the validity of our measure.

Participants

For this exploratory analysis, we included all children of the aforementioned samples where families filled out a short demographic questionnaire. This subsample consisted of 137 children, including 42 3-year-olds (mean = 43.04 months, SD = 3.25, range = 36 - 47, 23 girls), 46 4-year-olds (mean = 54.43 months, SD = 2.76, range = 48 - 59, 34 girls), and 49 5-year-olds (mean = 66.25 months, SD = 3.47, range = 60 - 71, 27 girls).

Procedure

Families of our kindergarten and online child sample were asked to fill out a brief demographic questionnaire. We asked for (1) the total number of household members, (2) the number of children, (3) age of the 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.

Analysis

To estimate the effects of social surrounding on gaze cue understanding, we fit GLMMs predicting the task performance by each of our questionnaire variables, controlling for age (in months, z-transformed), data collection mode (reference category: in-person supervised) and study version (reference category: hedge version). The models included random intercepts for each participant and each target position, and a random slope for symmetric target position within participants. Therefore, our null model closely resembled the structure from our first analysis (see Analysis section of *Does the balloon finding task induce variation?*; here: `performance ~ age + datacollection + studyversion + (symmetricPosition | subjID) + (1 | targetPosition)`). In order to combine data of our two study versions, we transformed continuous click responses from the hedge version into a discrete outcome. For the target position, we categorized two adjacent bins as one imaginary box. To measure participants' performance, we created imaginary box boundaries around the target's landing position and examined whether the participant's click response fell into this imaginary box. Across the two study versions, we could consequently model the participant's correct response (0/1) using a `Bernoulli` distribution with a logit link function. For model comparisons, we ran separate models, each with one of the following predictors as a fixed effects added to the null model: number of household members, number of children aged 0-18 in household, number of children aged 1-12 in household, hours spent in childcare each day, and age when subject entered childcare. In addition, we calculated three index scores. First, we calculated a sibling variety score according to Peterson (2000). Second, we implemented the modified version of Cassidy et al. (2005) (for more details, see Supplements). Third, based on our own data exploration, we calculated the amount of peer exposure determined as the number of siblings and the average hours spent in childcare (both z-transformed). We compared the models using WAIC (widely applicable information criterion) scores and weights (McElreath, 2020). As an indicator of out-of-sample predictive accuracy, lower WAIC scores stand for a better model fit. WAIC weights represent the probability that the model in question provides the best out-of-sample prediction compared to the other models.

Results

The model including our peer exposure index, as defined as the number of other children in the household and average hours spent in childcare, showed the best out-of-sample predictive accuracy. Note that we did not find a great difference in WAIC scores between the compared models (see Supplements for WAIC scores and weights). The model estimates were all considerably smaller than estimates of age, study version and data collection, and all 95% CIs included zero. For example, for our winning model, we found a peer exposure estimate of $\beta = 0.17$ (95% CI [-0.03; 0.36]), with the estimates of age being β =0.57 (95% CI [0.38; 0.77]), data collection mode being β =0.95 (95% CI [0.56; 1.35]), and study version β =1.87 (95% CI [0.25; 3.59]). Nevertheless, a general pattern emerges: exposure to a more variable social environment positively influenced children's gaze cue understanding. The number of people and, more specifically, children, as well as the more diverse their age, the more likely children were to understand the agent's gaze cue. The only predictor resulting in a negative estimate was the age at which a participant entered childcare, i.e., the later a child entered, the better performance in the task.

Discussion

We found that factors of children's social surrounding influenced their gaze cue understanding. Even though the effects are small and confidence intervals wide, it is remarkable that we were able to detect relationships between this fundamental social-cognitive ability and very distant, real life variables. Previous studies often focused on more complex, later developing social-cognitive abilities (e.g., false belief understanding). Apparently, systematic links between family factors and social-cognitive

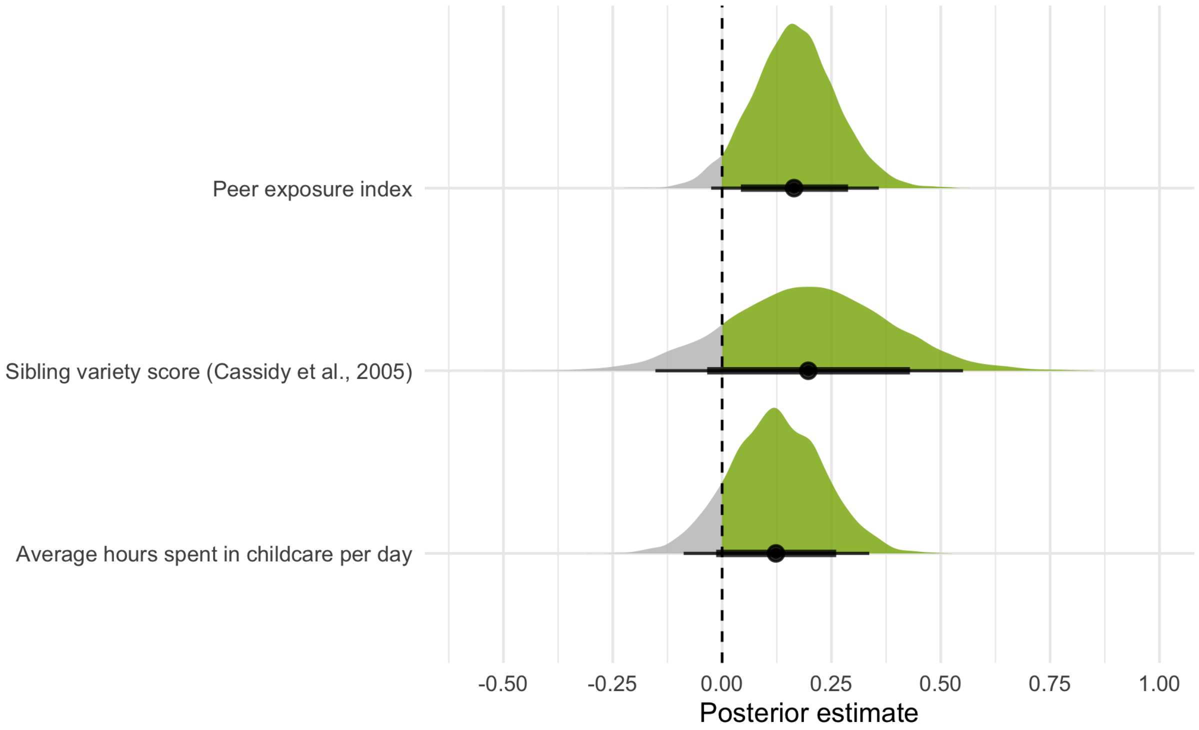

Figure 4. **External validity of the balloon finding task**. Factors of children's social surroundings and their influence on the probability of responding correctly. Models are ordered according to their WAIC scores, with the uppermost winning the model comparison. The graph shows the estimated density curves of a model's predictor coefficient. Only models performing better than our null model are included in the graph.

abilities can be found even when looking at more fundamental social-cognitive abilities like gaze cue understanding.

Discussion

We were able to show that our balloon finding task measures inter-individual variation between children and adults, alike. Our results suggest that the measured variation is systematic during the course of the same and different test days. Impressively, gaze cue understanding as measured by our task related to factors in children's everyday life experience.

Limitations

Future development / extending the task

Conclusion

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/).

Funding

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.

Ethics approval

Consent to participate

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

Consent for publication

Open access

Authors' contributions

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

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

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, *8*(6), 535–543. https://doi.org/10.1111/j.1467-7687.2005.00445.x

Bulgarelli, D., & Molina, P. (2016). Social Cognition in Preschoolers: Effects of Early Experience and Individual Differences. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01762

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395. https://doi.org/10.32614/RJ-2018-017

Cassidy, K. W., Fineberg, D. S., Brown, K., & Perkins, A. (2005). Theory of Mind May Be Contagious, but You Don't Catch It from Your Twin. *Child Development*, *76*(1), 97–106.

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

Devine, R. T., & Hughes, C. (2018). Family Correlates of False Belief Understanding in Early Childhood: A Meta-Analysis. *Child Development*, *89*(3), 971–987. https://doi.org/10.1111/cdev.12682

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.

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

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., & 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., & 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

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

McElreath, R. (2020). *Statistical rethinking: A Bayesian Course with Examples in R and Stan* (Second). Chapman and Hall/CRC.

Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., & Parlade, M. V. (2007 May-Jun). 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

Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of Mind Is Contagious: You Catch It from Your Sibs. *Child Development*, *65*(4), 1228–1238. https://doi.org/10.2307/1131316

Peterson, C. C. (2000). Kindred spirits: Influences of siblings' perspectives on theory of mind. *Cognitive Development*, *15*(4), 435–455. https://doi.org/10.1016/S0885-2014(01)00040-5

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.

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

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

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

Zhang, Z., Yu, H., Long, M., & Li, H. (2021). Worse Theory of Mind in Only-Children Compared to Children With Siblings and Its Intervention. *Frontiers in Psychology*, *12*, 5073. https://doi.org/10.3389/fpsyg.2021.754168

Supplements

Child sample

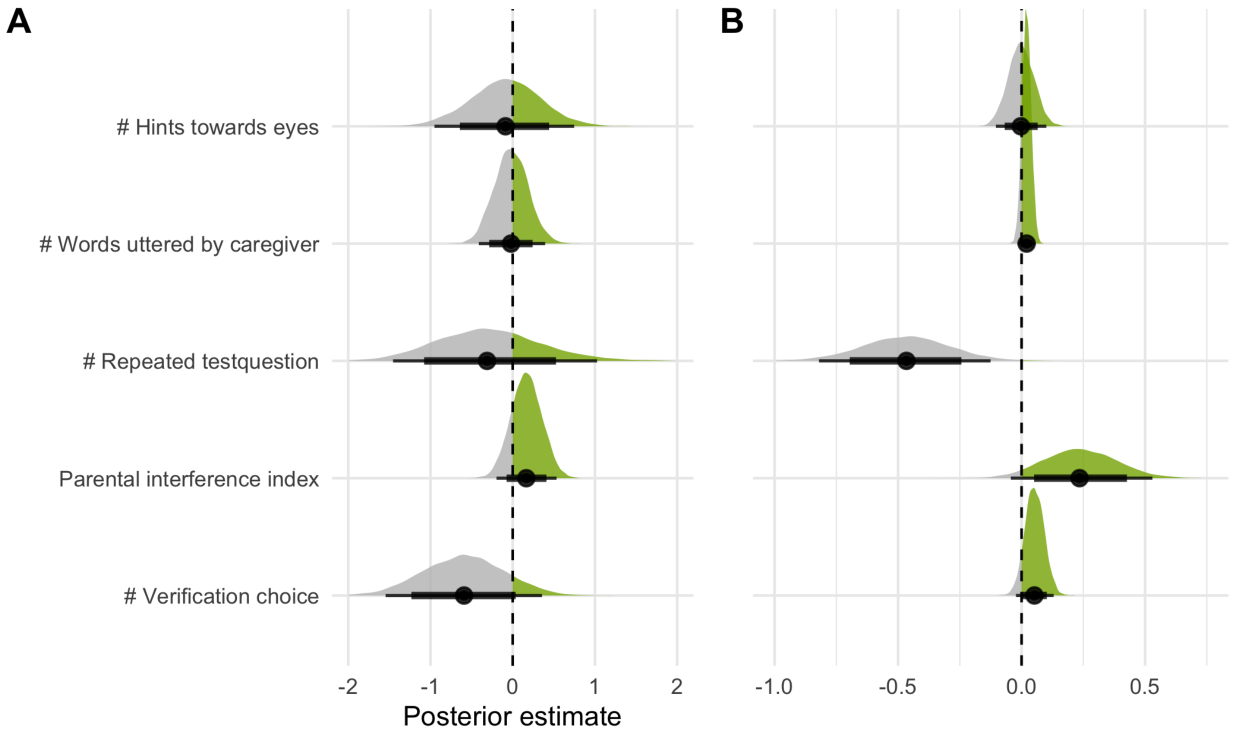

Figure 5. **Model comparison for exploratory webcam coding of parental interference**. Factors of parental interference and their influence on the probability of responding correctly. The graph shows the estimated density curves of a model's predictor coefficient. Models are ordered according to their 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.

Webcam coding. Comparing the performances of children across our two data collection modes, we found that children participating remotely were slightly more precise. This difference was especially prominent in younger participants 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. In an exploratory analysis, we coded parental behavior and environmental factors during remote unsupervised testing. 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* and *External Validity* sections, 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 **??**). Therefore, we conclude that the effect of parental interference was negligable and could, most likely, be explained as described above.

Appendix to external validity section.

Scoring of sibling variety scores. For assessing the external validity of our balloon finding task, we calculated two sibling variety scores based on the existing Theory of Mind literature. First, we followed the approach by Peterson (2000). Here, only-children as well as firstborns with siblings under one year scored 0 points; lastborns with siblings above 12 years scored 0.5 points; children with twins, firstborns with siblings over one year, and lastborns with at least one sibling under 13 years scored 1 point, middleborns with at least one older and younger sibling aged one to 12 years scored 2 points.

Second, we implemented the sibling variety score by Cassidy et al. (2005). The authors adjusted the original score of Peterson (2000) in the following way: only-children scored 0 points; children with a sibling under one year or above 12 years, and twins with no

Table 1

Model comparison for influences of children's social surrounding

Predictor	WAIC	SE_WAIC	Weight	ELPD_DIFF	SE_ELPD
Average hours spent in childcare per day	2,540.39	52.23	0.54	0.00	0.00
Peer exposure index	2,541.00	52.22	0.28	-0.30	0.92
# Children in household aged 0-18	2,541.65	52.30	0.02	-0.63	1.09
Sibling variety score (Cassidy et al., 2005)	2,541.76	52.33	0.01	-0.68	1.05
Sibling variety score (Peterson, 2000)	2,541.80	52.38	0.02	-0.71	1.10
Age of childcare entry	2,542.19	52.38	0.12	-0.90	1.32
# Children in household aged 1-12	2,542.26	52.34	0.00	-0.94	0.92
Null model	2,542.58	52.26	0.00	-1.09	0.79
# Household members	2,543.46	52.36	0.00	-1.54	0.97

Note. All models included random intercepts for each participant and each target position, and a random for symmetric target position within participants

other sibling scored 0.5 points; children with a sibling above one year or under 13 years scored 1 point; middleborns with at least one older and younger sibling aged one to 12 years scored 2 points. Twins with additional siblings scored depending on the age and number of their siblings.

The reasoning was that children between one and 13 years of age would engage in sibling play, while the youngest and most mature siblings would be less likely to participate in such. However, teenage siblings might provide opportunities for interesting discussions (Peterson, 2000).

WAIC scores and weights of the model comparison. As can be seen, ELPD difference scores are smaller than their respective standard errors. WAIC scores between models don't differ substantially (Table 1). All effects except when a child entered childcare positively influence performance.

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 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 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 *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 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 testdays, we reran the first testday with additional 50 participants. Additional seven subjects returned their submission, i.e., decided to leave the study early or withdrew their submission after study completion. Two additional participants 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 06:51min to complete the first part.

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 £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.

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