MAX PLANCK INSTITUTE
FOR EVOLUTIONARY ANTHROPOLOGY

LEUPHANA
UNIVERSITY LÜNEBURG

Rethinking Variation in Social Cognition: Gaze Following across Individuals, Ages, and Communities

Von der Fakultät Nachhaltigkeit
der Leuphana Universität Lüneburg zur Erlangung des Grades
Doktorin der Psychologie
– Dr. rer. nat. –

genehmigte Dissertation von

Julia Christin Prein
geboren am TT.MM.JJJJ in XXX

Eingereicht am: 1. Oktober 2024

Erstbetreuer: Prof. Dr. Manuel Bohn, *Leuphana Universität Lüneburg*
Zweitbetreuer: Prof. Dr. Sebastian Wallot, *Leuphana Universität Lüneburg*

Erstgutachter: Prof. Dr. Manuel Bohn, *Leuphana Universität Lüneburg*
Zweitgutachter: Prof. Dr. Sebastian Wallot, *Leuphana Universität Lüneburg*
Drittgutachter: Prof. Dr. Daniel Haun, *Universität Leipzig & Max-Planck-Institut für Evolutionäre Anthropologie*

Contents

Copyright Notice

 The individual contributions of this cumulative thesis have been or will be published (in chronological order) as follows:

Prein, J. C., Kalinke, S., Haun, D. B. M.*, & Bohn, M.* (2024). TANGO: A reliable, open-source, browser-based task to assess individual differences in gaze understanding in 3 to 5-year-old children and adults. *Behavior Research Methods, 56*(3), 2469–2485. https://doi.org/10.3758/s13428-023-02159-5

Prein, J. C., Maurits, L., Werwach, A., Haun, D. B. M.,* & Bohn, M.* (2024). *Variation in gaze understanding across the life span: A process-level perspective.* PsyArXiv. https://doi.org/10.31234/osf.io/dy73a

Bohn, M.*, Prein, J. C.*, Ayikoru, A., Bednarski, F. M., Dzabatou, A., Frank, M. C., Henderson, A. M. E., Isabella, J., Kalbitz, J., Kanngiesser, P., Keşşafoğlu, D., Koymen, B., Manrique-Hernandez, M., Magazi, S., Mújica-Manrique, L., Ohlendorf, J., Olaoba, D., Pieters, W., Pope-Caldwell, S., ... Haun, D. (2024). *A universal of human social cognition: Children from 17 communities process gaze in similar ways.* PsyArXiv. https://doi.org/10.31234/osf.io/z3ahv

Prein, J. C., Maurits, L., Werwach, A., Haun, D. B. M., & Bohn, M. (2024). *Variation in gaze following across the life span: A process-level perspective.* PsyArXiv. https://doi.org/10.31234/osf.io/dy73a

Further articles that were written in the context of this dissertation can be found in the Appendix and have been or will be published as follows (in chronological order):

Schuwerk, T., Kampis, D., Baillargeon, R., Biro, S., Bohn, M., Byers-Heinlein, K., Dörrenberg, S., Fisher, C., Franchin, L., Fulcher, T., Garbisch, I., Geraci, A., Grosse Wiesmann, C., Hamlin, K., Haun, D. B. M., Hepach, R., Hunnius, S., Hyde, D. C., Karman, P., ..., Prein, J., ... Rakoczy, H. (2021). *Action anticipation based on an agent's epistemic state in toddlers and adults. Child Development* (In-Principle Acceptance of Registered Report Stage 1: Study Design). PsyArXiv. https://doi.org/10.31234/osf.io/x4jbm

Bohn, M., Prein, J. C., Engicht, J., Haun, D., Gagarina, N., & Koch, T. (2023). *PREVIC: An adaptive parent report measure of expressive vocabulary in children between 3 and 8 years of age.* PsyArXiv. https://doi.org/10.31234/osf.io/hvncp

Bohn, M.*, Prein, J.*, Koch, T., Bee, R. M., Delikaya, B., Haun, D., & Gagarina, N. (2024). oREV: An item response theory-based open receptive vocabulary task for 3- to 8-year-old children. *Behavior Research Methods, 56*(3), 2595–2605. https://doi.org/10.3758/s13428-023-02169-3

Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A., Franchin, L., Garbisch, I., Grosse Wiesmann, C., Hamlin, J. K., Havron, N., Hay, J. F., Hermansen, T. K., Jakobsen, K. V., Kalinke, S., Ko, E.-S., Kulke, L., Mayor, J., Meristo, M., ..., Prein, J., ..., Schuwerk, T. (2024). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *Infancy, 29*(1), 31–55. https://doi.org/10.1111/infa.12564

* denotes shared first or last authorship.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Luctus venenatis lectus magna fringilla urna. Laoreet non curabitur gravida arcu ac tortor. Faucibus vitae aliquet nec ullamcorper sit amet risus nullam eget. Cras semper auctor neque vitae tempus quam pellentesque. Imperdiet massa tincidunt nunc pulvinar sapien et ligula ullamcorper. Et tortor at risus viverra adipiscing at. Congue nisi vitae suscipit tellus mauris. Habitant morbi tristique senectus et netus et malesuada fames ac. Eget mauris pharetra et ultrices neque. Aenean et tortor at risus viverra. Tempor orci dapibus ultrices in iaculis nunc sed augue. Euismod lacinia at quis risus sed vulputate odio ut enim. Id eu nisl nunc mi ipsum faucibus. Est lorem ipsum dolor sit amet. Eget velit aliquet sagittis id consectetur purus. Faucibus ornare suspendisse sed nisi. Pellentesque diam volutpat commodo sed egestas egestas fringilla.

Zusammenfassung

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Luctus venenatis lectus magna fringilla urna. Laoreet non curabitur gravida arcu ac tortor. Faucibus vitae aliquet nec ullamcorper sit amet risus nullam eget. Cras semper auctor neque vitae tempus quam pellentesque. Imperdiet massa tincidunt nunc pulvinar sapien et ligula ullamcorper. Et tortor at risus viverra adipiscing at. Congue nisi vitae suscipit tellus mauris. Habitant morbi tristique senectus et netus et malesuada fames ac. Eget mauris pharetra et ultrices neque. Aenean et tortor at risus viverra. Tempor orci dapibus ultrices in iaculis nunc sed augue. Euismod lacinia at quis risus sed vulputate odio ut enim. Id eu nisl nunc mi ipsum faucibus. Est lorem ipsum dolor sit amet. Eget velit aliquet sagittis id consectetur purus. Faucibus ornare suspendisse sed nisi. Pellentesque diam volutpat commodo sed egestas egestas fringilla.

Acknowledgements

1 Introduction

Please see Chapter 6. This will lead you to Study III.

1.1 How to include figures:

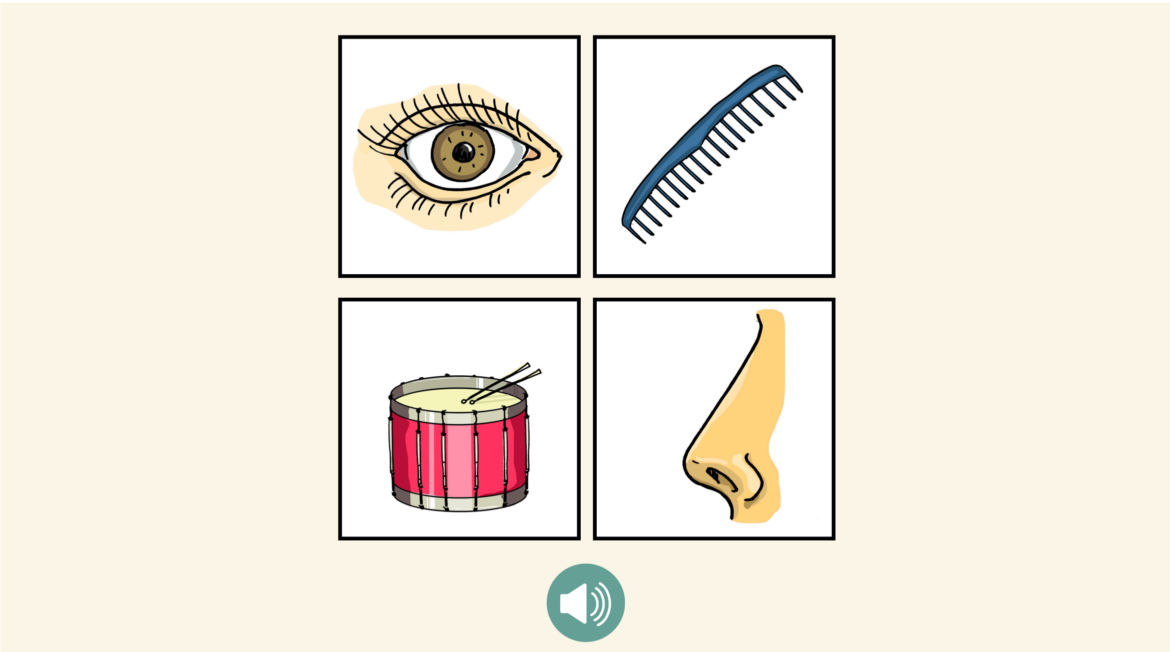

Figure 1.1: *OREV procedure**. Test for a figure caption.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Luctus venenatis lectus magna fringilla urna. Laoreet non curabitur gravida arcu ac tortor. Faucibus vitae aliquet nec ullamcorper sit amet risus nullam eget. Cras semper auctor neque vitae tempus quam pellentesque. Imperdiet massa tincidunt nunc pulvinar sapien et ligula ullamcorper. Et tortor at risus viverra adipiscing at. Congue nisi vitae suscipit tellus mauris. Habitant morbi tristique senectus et netus et malesuada fames ac. Eget mauris pharetra et ultrices neque. Aenean et tortor at risus viverra. Tempor orci dapibus ultrices in iaculis nunc sed augue. Euismod lacinia at quis risus sed vulputate odio ut enim. Id eu nisl nunc mi ipsum faucibus. Est lorem ipsum dolor sit amet. Eget velit aliquet sagittis id consectetur purus. Faucibus ornare suspendisse sed nisi. Pellentesque diam volutpat commodo sed egestas egestas fringilla.

Facilisi morbi tempus iaculis urna id volutpat. Faucibus in ornare quam viverra orci sagittis eu volutpat odio. Commodo elit at imperdiet dui. Condimentum mattis pellentesque id nibh tortor id aliquet lectus. Quis vel eros donec ac odio. Condimentum id venenatis a condimentum vitae sapien pellentesque habitant. Laoreet sit amet cursus sit amet dictum sit. Purus faucibus ornare suspendisse sed. Ornare arcu dui vivamus arcu. Scelerisque fermentum dui faucibus in. Nulla aliquet porttitor lacus luctus accumsan tortor posuere. Donec adipiscing tristique risus nec feugiat. Vitae turpis massa sed elementum tempus. Quis vel eros donec ac odio tempor. Iaculis eu non diam phasellus vestibulum lorem. Id porta nibh venenatis cras sed felis. Facilisi cras fermentum odio eu feugiat.

Lobortis feugiat vivamus at augue. Dignissim enim sit amet venenatis urna. Elementum sagittis vitae et leo duis ut diam quam. At lectus urna duis convallis convallis. Nisl condimentum id venenatis a condimentum vitae sapien pellentesque habitant. Mauris pellentesque pulvinar pellentesque habitant morbi tristique. Elit at imperdiet dui accumsan sit amet nulla facilisi morbi. Leo integer malesuada nunc vel risus commodo. Risus nec feugiat in fermentum posuere urna nec tincidunt. Quam id leo in vitae turpis massa. Duis ut diam quam nulla porttitor massa. Sem viverra aliquet eget sit amet tellus cras. Id nibh tortor id aliquet lectus proin nibh. Senectus et netus et malesuada fames ac. Leo urna molestie at elementum eu facilisis sed. Bibendum enim facilisis gravida neque convallis. Aenean vel elit scelerisque mauris pellentesque. Vitae purus faucibus ornare suspendisse.

Mauris nunc congue nisi vitae suscipit. Aenean euismod elementum nisi quis eleifend quam. Sed tempus urna et pharetra pharetra massa massa ultricies mi. Urna porttitor rhoncus dolor purus non. Rutrum tellus pellentesque eu tincidunt tortor. Fringilla urna porttitor rhoncus dolor purus non enim praesent. Tellus in hac habitasse platea dictumst vestibulum rhoncus est. Consectetur adipiscing elit pellentesque habitant. Et ligula ullamcorper malesuada proin libero. Viverra ipsum nunc aliquet bibendum enim.

In aliquam sem fringilla ut morbi tincidunt augue interdum. Rutrum quisque non tellus orci ac auctor augue mauris augue. Amet tellus cras adipiscing enim eu turpis egestas pretium. Nisl tincidunt eget nullam non nisi est sit amet facilisis. Diam quis enim lobortis scelerisque fermentum dui faucibus. Magna ac placerat vestibulum lectus mauris ultrices eros. Ac turpis egestas maecenas pharetra convallis posuere morbi leo. Velit aliquet sagittis id consectetur purus. Semper quis lectus nulla at volutpat diam. Orci nulla pellentesque dignissim enim sit amet venenatis urna. Lorem dolor sed viverra ipsum nunc. Risus nullam eget felis eget. Maecenas accumsan lacus vel facilisis. Sem integer vitae justo eget magna. Ac ut consequat semper viverra nam libero. Faucibus interdum posuere lorem ipsum dolor sit amet. Faucibus scelerisque eleifend donec pretium vulputate sapien nec. Leo vel fringilla est ullamcorper eget.

Euismod in pellentesque massa placerat. Volutpat diam ut venenatis tellus in metus vulputate eu scelerisque. Commodo quis imperdiet massa tincidunt nunc pulvinar sapien et. Vulputate dignissim suspendisse in est ante in nibh mauris. Proin fermentum leo vel orci porta non. Quis lectus nulla at volutpat diam ut venenatis tellus in. Leo vel orci porta non. Venenatis a condimentum vitae sapien pellentesque habitant morbi tristique. Quisque egestas diam in arcu. Vitae justo eget magna fermentum iaculis eu. Consequat ac felis donec et odio pellentesque diam. Nibh sed pulvinar proin gravida. Interdum consectetur libero id faucibus nisl tincidunt. Risus ultricies tristique nulla aliquet enim. Enim praesent elementum facilisis leo vel fringilla. Viverra nibh cras pulvinar mattis nunc.

Lorem dolor sed viverra ipsum nunc aliquet bibendum enim. Interdum consectetur libero id faucibus nisl tincidunt eget. Faucibus purus in massa tempor nec feugiat nisl pretium. Ac auctor augue mauris augue neque gravida in fermentum et.

In ante metus dictum at tempor commodo ullamcorper a lacus. Arcu vitae elementum curabitur vitae nunc sed. Eu scelerisque felis imperdiet proin fermentum leo. Orci ac auctor augue mauris augue neque gravida. Ac tortor dignissim convallis aenean et tortor at. Non arcu risus quis varius quam quisque. Ultricies integer quis auctor elit sed. Sit amet risus nullam eget felis eget nunc. Urna condimentum mattis pellentesque id nibh tortor id. Aliquam vestibulum morbi blandit cursus. Viverra tellus in hac habitasse platea dictumst vestibulum rhoncus est. Quis lectus nulla at volutpat diam. Lacus luctus accumsan tortor posuere. Non sodales neque sodales ut etiam sit amet nisl purus.

Cras semper auctor neque vitae tempus. Libero nunc consequat interdum varius. Orci ac auctor augue mauris augue neque gravida in. Ut placerat orci nulla pellentesque dignissim enim sit amet venenatis. Nunc mi ipsum faucibus vitae aliquet nec ullamcorper sit amet. Tortor dignissim convallis aenean et tortor at. Facilisi nullam vehicula ipsum a arcu cursus vitae. Felis imperdiet proin fermentum leo vel. Blandit volutpat maecenas volutpat blandit. Tortor id aliquet lectus proin. Enim eu turpis egestas pretium. Nunc faucibus a pellentesque sit amet porttitor. Fermentum iaculis eu non diam phasellus vestibulum. Porta lorem mollis aliquam ut porttitor. Amet nisl suscipit adipiscing bibendum. Malesuada fames ac turpis egestas integer eget aliquet nibh praesent. Libero volutpat sed cras ornare arcu. Arcu cursus euismod quis viverra nibh cras pulvinar mattis nunc. Lacinia quis vel eros donec ac odio tempor orci. Lacinia quis vel eros donec ac.

Varius duis at consectetur lorem donec massa sapien faucibus. Ac auctor augue mauris augue neque gravida in fermentum. Pellentesque habitant morbi tristique senectus et. Turpis egestas sed tempus urna et pharetra. Vitae auctor eu augue ut lectus. Est lorem ipsum dolor sit. Eget nunc lobortis mattis aliquam faucibus purus in massa. Donec ultrices tincidunt arcu non sodales neque. Sollicitudin nibh sit amet commodo nulla. Magna ac placerat vestibulum lectus mauris ultrices eros in. Sem viverra aliquet eget sit. Pellentesque massa placerat duis ultricies lacus sed turpis. Facilisi morbi tempus iaculis urna id volutpat lacus laoreet non. Massa massa ultricies mi quis. Aenean euismod elementum nisi quis eleifend quam adipiscing vitae. Ultrices eros in cursus turpis massa tincidunt dui. Augue lacus viverra vitae congue eu consequat. Eros in cursus turpis massa tincidunt dui. Facilisi morbi tempus iaculis urna id. Feugiat nibh sed pulvinar proin gravida hendrerit lectus a.

Mattis pellentesque id nibh tortor id aliquet lectus proin nibh. A cras semper auctor neque vitae tempus. Est pellentesque elit ullamcorper dignissim. Nulla facilisi etiam dignissim diam quis enim lobortis scelerisque fermentum. Laoreet sit amet cursus sit. Etiam tempor orci eu lobortis elementum nibh tellus molestie nunc. Nunc vel risus commodo viverra maecenas. Duis ut diam quam nulla porttitor. Ornare arcu odio ut sem nulla. Magna etiam tempor orci eu lobortis. Porttitor lacus luctus accumsan tortor posuere. Natoque penatibus et magnis dis parturient. Velit ut tortor pretium viverra suspendisse potenti nullam ac. Morbi leo urna molestie at. Nascetur ridiculus mus mauris vitae ultricies leo integer malesuada nunc. Sed viverra ipsum nunc aliquet bibendum enim facilisis gravida.

Risus ultricies tristique nulla aliquet enim tortor at auctor urna. Nisl rhoncus mattis rhoncus urna neque viverra justo nec ultrices. Id nibh tortor id aliquet lectus proin nibh nisl. Interdum varius sit amet mattis vulputate enim nulla aliquet porttitor. Ullamcorper eget nulla facilisi etiam. Scelerisque eleifend donec pretium vulputate sapien nec sagittis aliquam malesuada. Dui ut ornare lectus sit amet est placerat in egestas. Ultrices dui sapien eget mi proin. Sit amet nisl purus in mollis nunc. Senectus et netus et malesuada fames ac turpis. Nisi vitae suscipit tellus mauris a diam maecenas. Semper eget duis at tellus at urna condimentum.

Nibh nisl condimentum id venenatis. Ornare suspendisse sed nisi lacus sed viverra tellus in hac. Sed arcu non odio euismod lacinia. Sed libero enim sed faucibus turpis in eu mi. Dolor morbi non arcu risus quis varius quam quisque id. Etiam erat velit scelerisque in dictum non consectetur. Risus sed vulputate odio ut enim blandit volutpat maecenas volutpat. Libero volutpat sed cras ornare. Tellus cras adipiscing enim eu turpis egestas pretium aenean. Neque sodales ut etiam sit amet nisl.

A pellentesque sit amet porttitor eget dolor morbi. Libero volutpat sed cras ornare arcu dui vivamus arcu felis. Ultricies tristique nulla aliquet enim tortor at auctor. Tortor aliquam nulla facilisi cras fermentum odio eu feugiat. Natoque penatibus et magnis dis parturient montes nascetur. Ullamcorper eget nulla facilisi etiam dignissim diam quis enim. Sed elementum tempus egestas sed. Elementum eu facilisis sed odio morbi quis commodo. Proin fermentum leo vel orci. Id porta nibh venenatis cras sed felis eget velit. Vitae semper quis lectus nulla at volutpat.

Dictum fusce ut placerat orci nulla pellentesque dignissim enim. Turpis egestas sed tempus urna et pharetra pharetra massa. Pellentesque elit eget gravida cum sociis. Pellentesque id nibh tortor id aliquet lectus proin. Consectetur lorem donec massa sapien faucibus et molestie. Ut tellus elementum sagittis vitae et leo. Nisl condimentum id venenatis a condimentum vitae sapien pellentesque. A pellentesque sit amet porttitor eget dolor. In egestas erat imperdiet sed. Nec dui nunc mattis enim ut tellus elementum. Ut pharetra sit amet aliquam id diam maecenas ultricies mi. Massa enim nec dui nunc mattis enim. Nisl condimentum id venenatis a condimentum vitae sapien.

Habitasse platea dictumst vestibulum rhoncus est pellentesque. Tincidunt augue interdum velit euismod in pellentesque. Feugiat in fermentum posuere urna nec tincidunt. Tortor at auctor urna nunc id cursus metus aliquam. Pulvinar etiam non quam lacus. Imperdiet sed euismod nisi porta. Rhoncus urna neque viverra justo nec ultrices dui sapien eget. Ligula ullamcorper malesuada proin libero nunc consequat interdum varius sit. Mi ipsum faucibus vitae aliquet nec ullamcorper sit amet risus. Eu lobortis elementum nibh tellus molestie nunc. In est ante in nibh mauris cursus mattis molestie a. Sagittis id consectetur purus ut faucibus pulvinar elementum. Et netus et malesuada fames ac turpis egestas sed. Bibendum at varius vel pharetra vel turpis nunc eget lorem. Sem et tortor consequat id porta nibh venenatis cras sed. At urna condimentum mattis pellentesque.

1.2 Social Cognition

1.3 Gaze Following

1.4 Methodological Considerations

2 This Dissertation

2.1 Research Focus

2.2 Study Populations

2.3 Aims and Approaches

2.4 Ethics Statement

3 Results

4 General Discussion

5 Study I

Behavior Research Methods
https://doi.org/10.3758/s13428-023-02159-5

TANGO: A reliable, open-source, browser-based task to assess individual differences in gaze understanding in 3 to 5-year-old children and adults

Julia Christin Prein[1] · Steven Kalinke[1] · Daniel B. M. Haun[1] · Manuel Bohn[1,2]

Accepted: 2 June 2023

Abstract
Traditional measures of social cognition used in developmental research often lack satisfactory psychometric properties and are not designed to capture variation *between* individuals. Here, we present the TANGO (Task for Assessing iNdividual differences in Gaze understanding-Open); a brief (approx. 5–10min), reliable, open-source task to quantify individual differences in the understanding of gaze cues. Localizing the attentional focus of an agent is crucial in inferring their mental states, building common ground, and thus, supporting cooperation. Our interactive browser-based task works across devices and enables 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. Our task measures inter-individual differences in a child ($N = 387$) and an adult ($N = 236$) 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. Associations with social-environmental factors and language skills speak to the validity of the task. This work shows a promising way forward in studying individual differences in social cognition and will help us explore the structure and development of our core social-cognitive processes in greater detail.

Keywords Social cognition · Individual differences · Gaze cues · Cognitive development

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 et al., 2012; Rakoczy, 2022; Wellman et al., 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 designing interventions (Happé et al., 2017; Kidd et al., 2018; Lecce et al., 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). The most common, recurring research questions are concerned with the developmental sequence of social-cognitive abilities (e.g., Wellman & Liu, 2004), and which factors drive the development of social cognition (Devine & Hughes, 2018; Gola, 2012). For example, Okumura and colleagues asked how early gaze-following and object processing relate to later language development (Okumura et al., 2017). In general, individual differences studies often focus on the relationship between social-cognitive abilities and: (1) family influences, (2) other cognitive constructs, and (3) social behavioral outcomes (for an overview, see Slaughter and Repacholi, 2003). Studies

Haun, Daniel B. M. and Bohn, Manuel contributed equally to this work.

✉ Julia Christin Prein
julia_prein@eva.mpg.de

[1] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

[2] Institute of Psychology, Leuphana University Lüneburg, Lüneburg, Germany

Published online: 10 July 2023

on social-cognitive abilities and family influences include the effect of parenting practices (for a review, see Pavarini et al., 2013), attachment quality (e.g., Astor et al., 2020), mental state talk (Gola, 2012; Hughes et al., 2011; Lecce et al., 2014), and family background as parental education, occupation, sibling interaction and childcare (Bulgarelli & Molina, 2016; Cutting & Dunn, 1999; Dunn et al., 1991). Another group of individual differences studies focuses on the interplay of social and physical cognition (Herrmann et al., 2010), executive functions (Benson et al., 2013; Buttelmann et al., 2022; Carlson & Moses, 2001; Carlson et al., 2004; Hughes & Ensor, 2007), and language abilities (McEwen et al., 2007; Milligan et al., 2007; Okumura et al., 2017). Studies on social behavioral outcomes measured the interplay of social cognition and prosociality (for a review, see Imuta et al., 2016; Walker, 2005), stereotypes, resource allocations (Rizzo & Killen, 2018), and moral intentions (Sodian et al., 2016).

However, developmental psychologists are frequently 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., Poulin-Dubois et al., 2023; Sodian, 2023; 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 et al., 2020) or – when they did – showed mixed results on test–retest reliability (Hughes et al., 2000; Mayes et al., 1996).

To give an example: the most commonly applied prototypical measure for social cognition is the change-of-location false belief task (Baron-Cohen et al., 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).

However, from this age onwards, the change-of-location task shows ceiling effects and has very limited diagnostic value (Repacholi, 2003). Thus, this task seems well suited to track a particular group-level developmental transition, yet it fails to capture individual differences (cf. "reliability paradox," Hedge et al., 2018). As Wellman (2012) put it, "it's really only passing/failing one sort of understanding averaged across age" (p. 317). This has profound implications for what studies on individual differences using this task (or others) can show. Poor measurement of social cognition on an individual level is likely to conceal relations between different aspects of cognition and may obscure developmental change. For example, Sodian et al. (2016) neither found a correlation between two moral Theory of Mind False Belief and Intention tasks at 60 months, nor a relationship between these two factors and implicit False Belief understanding at 18 months.

The "Sandbox task" is one of the few tasks that attempt to overcome these methodological challenges (Begeer et al., 2012; Bernstein et al., 2011; Coburn et al., 2015; Mahy et al., 2017; Sommerville et al., 2013). This continuous FB task measures the degree to which the estimate of another's belief is biased by one's own knowledge. Recent work questions the interpretation of this measure (Samuel et al., 2018a, b): it is unclear whether a smaller egocentric bias can be directly translated into a better mental state reasoning ability. Another evaluation criterion should, therefore, be whether a task captures meaningful variability in performance; that is, differences in test scores should correspond to differences in the social-cognitive ability in question.

Thus, developmental psychology faces a dilemma: many research questions rely on measuring individuals' development, yet, there is a lack of tasks to measure these individual differences reliably. To capture the emergence of social-cognitive abilities and their relation to social factors in greater precision and detail, we must consequently address the methodological limitations of existing study designs (Hughes et al., 2011; Hughes & Leekam, 2004).

Schaafsma et al., (2015) compiled a "wish list" for new social-cognitive paradigms. They advocated for parametric – instead of dichotomous – measures covering proficiency as a range, avoiding floor and ceiling effects, and showing satisfactory test–retest reliability estimates (see also Beaudoin et al., 2020; Hughes & Devine, 2015). New tasks should capture variation across age groups, including older children and adults (Repacholi and Slaughter, 2003). Another goal in creating new tasks should be to focus on the "face value": measures should probe the underlying social-cognitive ability as straight-forward and directly as possible. Keeping task demands minimal is also beneficial for using the paradigm in a variety of different cultural, clinical, and demographic contexts (Molleman et al., 2019). The task should serve as a proxy for behavior as it appears in the real world and should be validated in relation to real-world experiences (Repacholi and Slaughter, 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, reliable, and valid 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 et al., 2021; Byers-Heinlein et al., 2021; Coelho et al., 2006; Del Bianco et al., 2019; Frischen et al., 2007; Hernik & Broesch, 2019; Itakura & Tanaka, 1998; Lee et al., 1998; Moore, 2008; Shepherd, 2010; Tomasello et al., 2007). 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.

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). Gaze is integral for many more sophisticated social-cognitive abilities, for example, inferences about knowledge states. As such, the eyes have been regarded as a "window into the mind" (Shepherd, 2010). Monitoring another's attention also supports building a common ground, which is important for action coordination and cooperative social interactions (Bohn & Köymen, 2018; Tomasello et al., 2007). In addition, gaze and language development seem to be related (Brooks & Meltzoff, 2005). Gaze facilitates word learning by helping to identify the referent of a new word and has been regarded as a crucial signal of nonverbal communication (Hernik & Broesch, 2019; Macdonald & Tatler, 2013).

While the emergence of gaze following has been well established, less is known about the developmental trajectory throughout childhood and adolescence. One possibility is that our social-cognitive ability in question is fully developed once emerged in infancy. However, many cognitive abilities continue to develop beyond early childhood (e.g., Gathercole et al., 2004 for working memory; Raviv & Arnon, 2018 for visual statistical learning). Therefore, children could potentially improve in understanding gaze, fine-tuning the performance of the already existing skill. Consequently, we aimed to assess the differentiation of the ability to understand gaze. Our goal was *not* to establish the youngest age at which children understand gaze cues. Rather, we wanted to examine how that ability changes with age. To accurately measure developmental change, we were interested in capturing individual variability.

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 is 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 web application. Previous research has successfully used online study implementations that compare well to in-person data collection (Bohn et al., 2021a, b; Frank et al., 2016). 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, 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/tango-demo), and a live demo version can be found under: https://ccp-odc.eva.mpg.de/tango-demo/.

The web app was developed using 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 viewport 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 the 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 metadata, 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. Personalized links can be created by passing on URL parameters.

Stimuli

Our newly implemented task asks children and adults to search for a balloon. The events proceed as follows (see Fig. 1B and C). An animated agent (a sheep, monkey, or

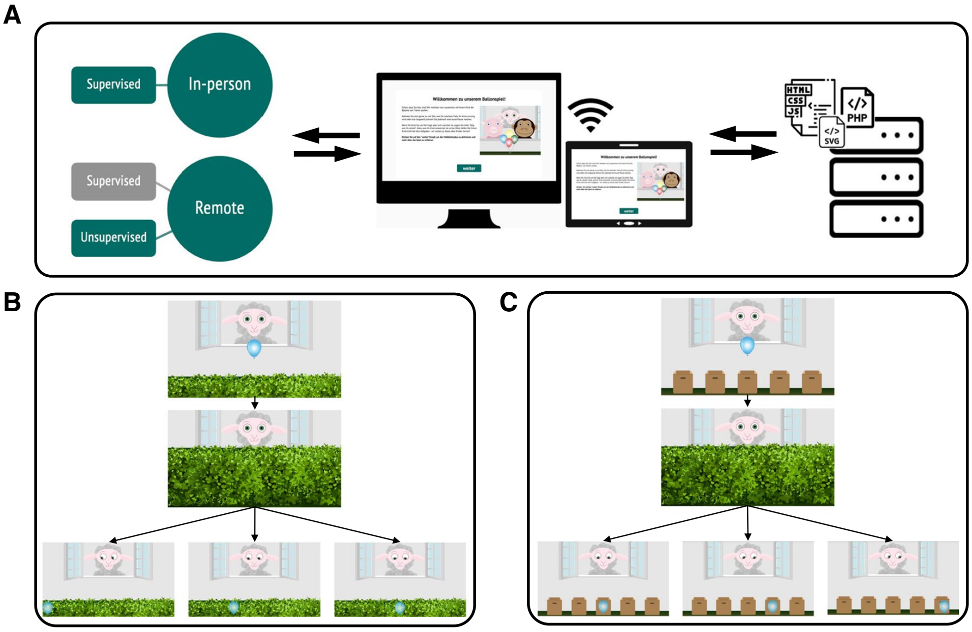

Fig. 1 Study setup. **A** Infrastructure for online testing. (i) Subjects aged 3 to 99+ can participate. Data collection can take place anywhere: online, in kindergartens, or in research labs. (ii) The task is presented as a website that works across devices. (iii) The scripts for the website and the recorded data are stored on secure in-house servers. **B** Hedge version (continuous) of the TANGO. (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. Three exemplary target locations are shown to depict how indicative the agent's gaze cues are in determining the target's location. The transparent target is only shown for an illustrative purpose (not visible during the test). **C** Box version (discrete) of the TANGO. Number of boxes (min. 1; max. 8) as potential hiding locations can be set according to the researcher's need

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: the pupils and iris move so 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 s. The landing of the target is accompanied by a tapping sound. Once the target landed, the instructor's voice asked "Where is the balloon?". To confirm the participant's click, a short plop sound plays, and a small orange circle appears at the location of choice. Participants do not receive differential feedback so that learning effects are reduced, and trials stay comparable across the sample. If no response is registered within 5 s 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 touching 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 Supplements 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 min.

Study versions

We designed two study versions that differ in the target's final hiding place and, consequently, in 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 can choose 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 Fig. 1B and C).

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 the TANGO captures inter-individual variation in a child and adult sample. Furthermore, we were interested in whether and how the data collection mode (in-person vs. remote) influences responses. Since we expected a greater difference in responses between the two data collection modes for children, the analysis of data collection mode was restricted to a child sample.

Task design, data collection, and sample sizes were pre-registered: https://osf.io/snju6 (child sample) and https://osf.io/r3bhn (adult sample). The analyses reported here were not pre-registered but followed the structure of the ones specified in the above pre-registrations (see Footnotes for deviations). The additional analyses mentioned in the pre-registrations (e.g., computational model) address separate research questions (e.g., process-level account of gaze understanding) and will be reported elsewhere. In this paper, we focus on the methodological and psychometric aspects of our task.

The study design and procedure obtained ethical clearance by the MPG Ethics commission Munich, Germany, falling under a packaged ethics application (Appl. No. 2021_45), and was approved by an internal ethics committee at the Max Planck Institute for Evolutionary Anthropology. The research adheres to the legal requirements of psychological research with children in Germany.

Participants were equally distributed across the two study versions. Data were 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).

We pre-registered the replacement for participants that finished fewer than four test trials. This was not the case for any participant. One child stopped participation after 12 test trials but was included in the sample due to the pre-registered replacement rule. Two additional participants were recruited but not included in the study because the participant did not feel comfortable interacting with the tablet alone (n = 1), or due to an originally miscalculated age of the child (n = 1).

For our remote child sample, we recruited families via an internal database of children living in Leipzig, Germany, whose parents volunteered to participate in child development studies and who indicated an interest in online studies. Families received an email with a short study description and a personalized link. If they had not participated in the study within 2 weeks, they received a reminder via e-mail. The response rate to invitations after the reminder was ~ 50%.

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). Of these, three families participated twice. In these cases, we only kept the data sets from the first participation.

Four additional participants were recruited but not included in the study because they were already part of the in-person kindergarten sample (n = 3), or because of unknown age (n = 1).

Please note that we did not collect participant-specific demographics. In the following, we aim to provide context and generalizations based on the broader community and the larger pool of potential participants. Children in our sample grow up in an industrialized, urban Central-European context in a city with approximately 600,000 inhabitants. They often live in nuclear two-generational families with few household members. 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. The median individual monthly net income in the year 2021 was ~ 1,600€ for the city of Leipzig.

Adults were recruited via *Prolific* (Palan & Schitter, 2018). *Prolific* is an online participant recruitment service with a predominantly European and US–American subject pool. One hundred English speakers with an average age of 31.34 years (SD = 10.77, range = 18–63, 64 females) were included. Participants live in a variety of different countries: the UK, Italy, Spain, Poland, Netherlands, Canada, Australia, Ireland, South Africa, Norway, Portugal, France, Austria, Finland, Greece, Germany, the U.S., Mexico, Chile, Iceland, New Zealand, Czech Republic, Hungary, Latvia, and Switzerland. In this sample, most participants resided in the United Kingdom (n = 47), South Africa (n = 8), and Portugal (n = 6). Additional detailed information can be found in the data set online. For completing the study, subjects were paid above the fixed minimum wage (on average £10.00 per hour; see Supplements for further detail).

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. At the beginning of the online study, families were invited to enter our "virtual institute". We welcomed them with a short introductory video of the study leader, 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 on how to set up 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 children themselves generated the answers. 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 a touchscreen, the hedge version was shown. Here, children could directly click on the touchscreen to indicate where the target is. For families that used a computer without a touchscreen, the website presented the box version of the task. We assumed that younger children in our sample would not be acquainted with using 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 descriptions were included in our analyses. We ran all analyses in R version 4.3.0 (2023-04-21) (R Core Team, 2022). Regression models were fitted 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 symmetric target position, and a random slope for symmetric target position within participants (model notation in R: performance ~ age + datacollection + symmetricPosition + trialNr + (1 + symmetricPosition + trialNr | subjID)).[1]

[1] In the pre-registration (https://osf.io/snju6), we specified the following model structure: "All models will include a fixed effect of target centrality and age, a random intercept for ID and a random slope for trial number by ID. For both study versions, we will compare the above specified null models with a model including data source (live vs. online) as a fixed effect." Or, in R model formula: R: performance ~ target_centrality + age + (1 | ID) + (1 + trial | ID). In this paper, we added symmetricPosition (synonymous to target_centrality) as a random slope because we expected that this item effect could vary between participants. To be better able to interpret trial

Here, 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 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% credible interval (CrI)) for the age and data collection estimates.

Results

Children showed nearly perfect precision in the first training trial. As visual access to the target location decreased in the subsequent training trials, imprecision levels increased (see Supplements). Within test trials, children's imprecision levels did not vary as a function of trial number. We take this as evidence that (A) children were comfortable touching the screen, (B) children understood the task instructions insofar as they aimed at locating the target, and (C) our experimental design successfully manipulated task difficulty.

We found a strong developmental effect: with increasing age, participants got 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 Fig. 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 3-year-olds were more precise in their responses than some 5-year-olds. Even though variation is smaller, we could even find inter-individual differences in the adult sample.

As Fig. 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 was 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.[2]

Our GLMM analysis corroborated the visual inspection of the data: in the hedge version, the estimates for age (β = – 0.32; 95% CrI [– 0.41; – 0.23]) and data collection mode – 0.31 (95% CrI [– 0.48; – 0.14]) were negative and reliably different from zero. In the box version, the estimate of age (β = 0.68 (95% CrI [0.44; 0.93]) and the estimate of data collection mode (β = 1.10 (95% CrI [0.66; 1.56]) 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 that our remote sample performed more accurately than our in-person sample.

There was no effect of trial number (hedge version: β = 0.00; 95% CrI [– 0.02; 0.01]; box version: β = – 0.02; 95% CrI [– 0.05; 0.01). However, trials differed in difficulty depending on where the target landed (hedge version: β = 0.47; 95% CrI [0.40; 0.54]; box version: β = – 1.59; 95% CrI [– 1.88; – 1.31). When the target landed closer to the center of the screen, participants were more accurate in locating it.

Discussion

Our task measured inter-individual differences in both children and adults; that is, we found substantial variation in individuals across age groups. For example, some 3-year-olds showed greater precision levels than some 5-year-olds. This holds across both study versions. However, due to the continuous study design, the hedge version was able to capture more fine-grained differences in individual performance. We see substantial developmental gains: with increasing age, participants became on average more and more precise in locating the target. The 5-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 web app with cartoon-like features kept children interested 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 design.

Footnote 1 (continued)

number effects, we decided to include it as a fixed effect. Data collection mode (formerly named "data source") proved as a meaningful predictor and was accordingly added to the model.

[2] 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 3-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.

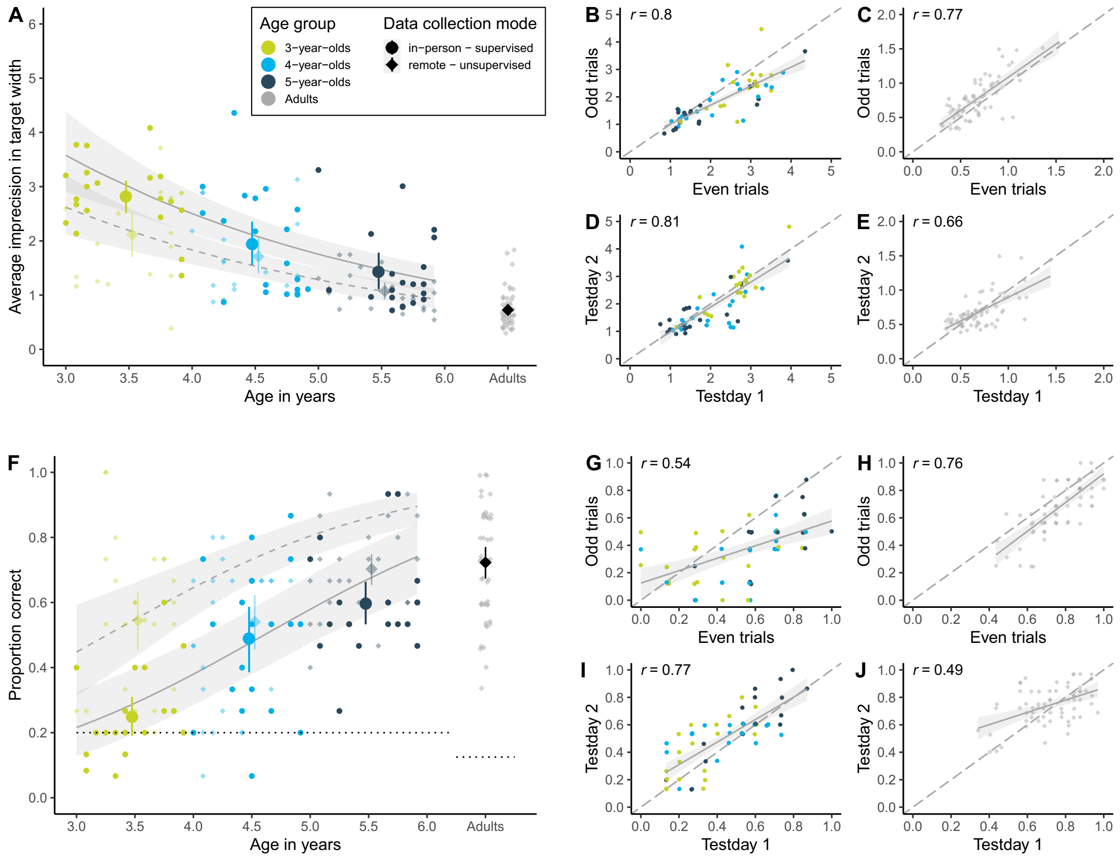

Fig. 2 Measuring inter-individual variation. **A** Developmental trajectory in the continuous hedge version. Performance is measured as imprecision, i.e., the absolute distance between the target's center and the participant's click (averaged across trials). The unit of imprecision is counted in the width of the target, i.e., a participant with imprecision of 1 clicked on 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 the discrete box version. Performance is measured as the proportion of correct responses, i.e., how many times the participant clicked on the box that contained the target. The *dotted black line* shows the level of performance expected by chance (for child sample 20%, i.e., one out of five boxes; for adult sample 12.5%, i.e., one out of eight 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. *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. For adult data in (A) and (F), we added minimal horizontal and vertical noise to avoid overplotting. The *shape of data points* represents data collection mode: *opaque circles* for in-person supervised data collection and *translucent diamonds* for remote unsupervised data collection. The *color of data points* denotes age group. 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*

Internal consistency and test–retest reliability

As a next step, we aimed to investigate whether the variation that we captured with the TANGO is reliable. We assessed internal consistency (as split-half reliability) and test–retest reliability. Task procedure, data collection, and sample sizes were pre-registered (https://osf.io/xqm73 for the child sample and https://osf.io/nu62m for the adult sample). Participants were equally distributed across the two study versions. Data was collected between July 2021 and June 2022.

The study design and procedure obtained ethical clearance by the MPG Ethics commission Munich, Germany, falling under a packaged ethics application (Appl.

No. 2021_45), and was approved by an internal ethics committee at the Max Planck Institute for Evolutionary Anthropology. The research adheres to the legal requirements of psychological research with children in Germany.

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).

Additional 65 children were recruited but not included in the analysis due to absence on the second test day ($n = 49$), canceled testing because of COVID-19 cases in the kindergarten ($n = 7$), children did not want to participate a second time ($n = 5$), children already participated in the first data collection round and were included in the above-mentioned *Individual Differences* sample ($n = 3$), or children did not understand the task instructions ($n = 1$; manifested in too early clicking in the training trials while the instructions were still playing, and no clicking by themselves in the test trials). Two additional children were recruited for the first day (as backup) in case another child would be absent on the second test day. Similar to our first study, we did not collect participant-specific demographics. For a community-based description of our participant pool, see Participant section of the first study.

As in our first study, adults were recruited via *Prolific* (Palan & Schitter, 2018). The adult sample included 136 English speakers with an average age of 25.73 years (SD = 8.09, range = 18–71, 87 females; see Supplements for further details). Most participants resided in South Africa ($n = 48$), the United Kingdom ($n = 19$), and the United States ($n = 14$). See Supplements and the available online data set for more detailed information.

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, were 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 bins 1 & 2, bins 3 & 4, etc.). In the box version, we ensured that each of the five boxes occurred exactly three times during test trials. Adults in the hedge version received 30 test trials, each of the ten bins occurring exactly three times. Adults in the box version received 32 test trials, with each of the eight boxes occurring exactly four times. For the four training trials, we repeated a fixed order of random bins/boxes. For the adult sample, we decided to increase the number of trials in order to get more accurate reliability estimates. Trial numbers were multipliers of the possible target locations and therefore differed between hedge and box versions. For the child sample, we stuck to the same number of trials to not risk higher attrition rates.

Analysis

We assessed reliability in two ways. First, we focused on internal consistency by calculating split-half reliability coefficients.[3] For each subject, trials were split into odd and even trials. Performance was aggregated and then correlated using *Pearson* correlation coefficients. For this, we used the data of the first test day. Performance was defined according to each 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., (2022) 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 split-half 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 test–retest reliability. We calculated performance scores (depending on the 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

[3] The assessment of internal consistency was not pre-registered and was included as an additional measure of reliability.

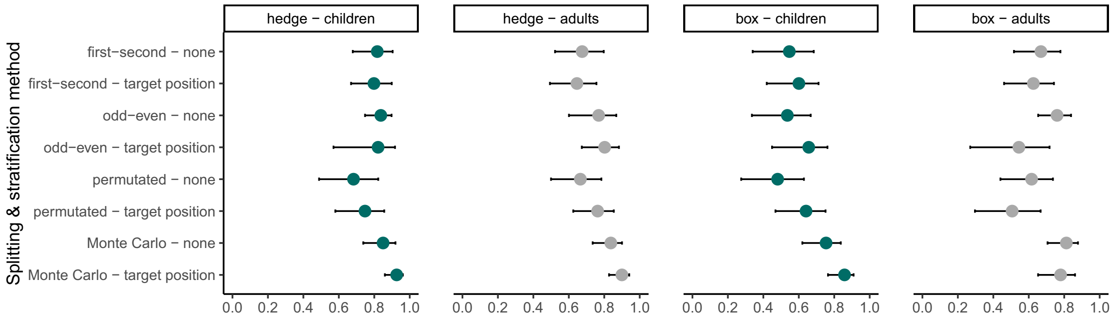

Fig. 3 Internal consistency. Reliability coefficients per splitting method, stratification level, study version, and age group. *Error bars* show the 95% confidence intervals of the coefficient estimates, calculated with the function by_split from the splithalfr package (Pronk et al., 2021)

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 indicate a high association between test days.

Results

We found that the TANGO measured systematic variation: split-half and test–retest reliability was medium to high. For internal consistency, we show traditional odd-even splits on our data and the corresponding *Pearson* correlation coefficients in Fig. 2B, C, G, and H.

Figure 3 compares split-half reliability coefficients by splitting and stratification method (Pronk et al., 2021). In the hedge version, the split-half reliability coefficients ranged from 0.65 to 0.93. In the box version, split-half reliability coefficients ranged from 0.48 to 0.86. Similar 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 most cases, stratifying by target position led to similar or even higher estimates compared to no stratification. As expected, we found higher coefficients for the samples with higher variation, i.e., for our continuous hedge version of the task.

For test–retest reliability, we show the association between raw performance scores of the two test days and corresponding *Pearson* correlation coefficients in Fig. 2D, E, I and J.[4] See Supplements for reliability estimates by age group.

The age-corrected, GLMM-based retest reliabilities for children yielded similar results. In the hedge version, the correlation between test days was 0.89 (95% CrI [0.64;1.00]). In the box version, the correlation between test days was 0.91 (95% CrI [0.70;1.00]).

For both study versions, reliability estimates based on the GLMM approach were higher than the *Pearson* correlations. The GLMM-based estimates are less noisy due to the fact that the model uses all available information (e.g., participant age) and does not rely on data aggregation across trials.

Discussion

Our results indicated that the measured variation was systematic. As 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 saw in Fig. 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.

[4] 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 $r = 0.88$. Including the two adults yielded a *Pearson* correlation coefficient of $r = 0.73$.

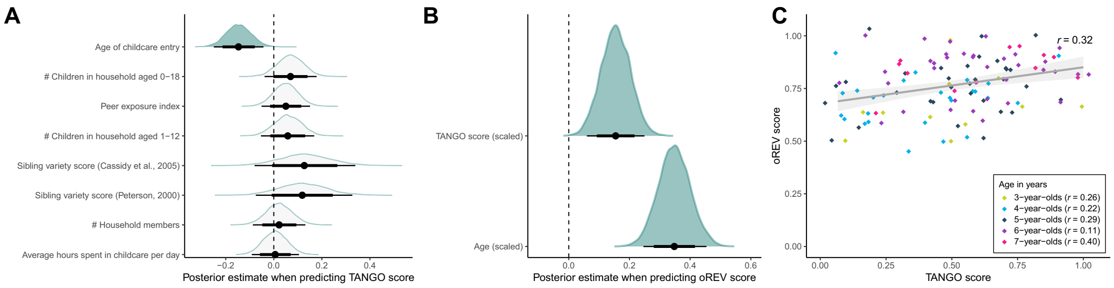

Fig. 4 Validity of the TANGO. **A** Influence of social-environmental factors on gaze understanding. **B** Influence of gaze understanding on receptive vocabulary. For (A) and (B), the graphs show the posterior distributions for the respective predictor of each model. *Filled green density curves* show that adding the respective predictor improved the model fit compared to the null model. *Black dots* represent means, *thicker black lines* 80% CrI and *thinner black lines* 95% CrI. The oREV score is the proportion of correctly selected pictures in the receptive vocabulary task. Similarly, the TANGO score refers to proportion of correctly located targets (see Supplements for further detail). **C** Influence of gaze understanding on receptive vocabulary by age. The *regression line* with 95% CI shows a smooth conditional mean based on a generalized linear model, with *Pearson*'s correlation coefficient *r*. *Dots* show the mean performance for each subject averaged across trials with minimal horizontal and vertical noise added to avoid overplotting. The *color of dots* denotes age group

Validity

After having probed our new testing infrastructure and the psychometric properties of the TANGO, we aimed at establishing its validity. One way to assess validity is to correlate the social-cognitive ability in question to concepts that are thought to be theoretically related. Social cognition is often described as developing in response to social interaction (Devine & Hughes, 2018; Hughes & Leekam, 2004). It is assumed that opportunities to play, communicate and argue with peers help children to understand the human mind. Therefore, many studies link social cognition to opportunities for social interaction captured in demographic variables such as parent–child interaction quality and quantity, mental state talk, and center-based childcare (Bulgarelli & Molina, 2016; Dunn et al., 1991; Pavarini et al., 2013). In particular, family constellation, the number and age of siblings, and their interaction have been linked to social cognition (Cassidy et al., 2005; Dunn et al., 1991; Perner et al., 1994; Peterson, 2000; Zhang et al., 2021).

To assess such external validity for the TANGO, we handed out a brief demographic questionnaire to families of our kindergarten and online child sample and 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 daycare, and if yes, (5) since when and (6) for how long on an average day. 109 families filled out the questionnaire and were included in the analysis. We used parents' responses to construct different scores suggested in the literature (Cassidy et al., 2005; Peterson, 2000), capturing aspects of children's opportunities for social interaction with adults and peers. Only the predictor "age of childcare entry" improved the model fit compared to the null model (see Fig. 4A; for model comparisons, see Supplements): the older the children were when entering childcare, the less likely they were to correctly use the available gaze cue. Figure 4A shows that all other predictor scores were positively linked to gaze understanding. Effect sizes were probably influenced by the lack of variance in the predictors: variables like household size and number of siblings typically vary very little among German households (see Supplements for distribution characteristics of the predictors). Albeit the effects were weak, they are consistent with the literature.

In addition, children's sensitivity to gaze has been linked to language acquisition (Brooks & Meltzoff, 2005; Del Bianco et al., 2019; Okumura et al., 2017). Discovering the attentional focus of your counterpart is thought to facilitate word learning, for example by identifying the referent of a new word (Tomasello, 2003). For 117 children, we also collected data with a receptive vocabulary test (oREV; Bohn et al., 2023) approximately 6 months (mean = 0.52 years, SD = 0.08, range = 0.06–0.80) after their participation in the TANGO. In the oREV task, children are shown four pictures (see Supplements for further detail) and hear a verbal prompt asking them to select one of the pictures. The oREV score is the proportion of correctly selected pictures. We found a substantial relationship between gaze understanding 6 months prior and receptive vocabulary, even when correcting for age (see Fig. 4B and C). Taken together, our newly developed task shows connections to external variables and psychological constructs that are characteristic of measures of social cognition.

General discussion

We have presented a new experimental paradigm to study gaze understanding across the lifespan. This paper contributes to methodological advances in developmental psychology in the following ways: first, we captured fine-grained individual differences in gaze understanding at different ages – from early childhood until adulthood. Individuals behaved consistently differently from one another (i.e., we found substantial variation between individuals across age groups). Second, our task showed satisfactory psychometric properties with respect to internal consistency and test–retest reliability estimates. Third, our new browser-based testing infrastructure ensures standardized, portable data collection at scale, both remotely as well as in person. In sum, the TANGO provides a step toward more robust and reliable research methods, especially with regard to measuring developmental change in a fundamental social-cognitive ability. The web app (https://ccp-odc.eva.mpg.de/tango-demo/) and its source code (https://github.com/ccp-eva/tango-demo) are freely accessible for use and modification.

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) toward 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, revealing relations between different aspects of cognition and children's real-life social surroundings, and for meaningful comparisons across human cultures and across animal species. 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 than the categorical box version. Both study versions exhibit high test–retest reliability, also when controlling for age. Therefore, when a sufficient number of trials is presented, the box version of the task can also yield reliable individual estimates (cf. Hughes et al. (2000); improved reliability through aggregation). When testing time is limited (and the number of trials might be low), we recommend using the continuous study version for higher internal consistency. However, the categorical box version demonstrates design features that might be preferable in some research contexts: for example, 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).

In the split-half reliability calculations, the more accurately the statistical method represents the task structure, the higher the reliability estimates are. Therefore, we argue that future research should aim at implementing statistical analyses that mirror the complexity of the experimental design. Theoretically informed, computational cognitive models are a promising approach forward (Haines et al., 2020). Computational models 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 et al., 2023).

The TANGO fulfills several demands that were proposed by Schaafsma et al. (2015)'s wish list: it measures proficiency on a continuum, avoids floor and ceiling effects, measures variation across age ranges, shows satisfactory reliability estimates, and has a high face value.

In addition to the new task design itself, we designed a new testing infrastructure. The TANGO 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., target positions) on a trial-by-trial basis. Presented agents, voice-over instructions, and objects can be easily adapted for future task modifications or specific linguistic and cultural settings.

The browser-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 and supervised in kindergartens with children who participated remotely at home. Our results suggest a comparable developmental trajectory of gaze understanding in both samples. Children in the remote sample were slightly more precise. This effect was most pronounced in the 3-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 versatile framework for addressing further research questions on social-cognitive development.

With respect to validity, we found that performance in the TANGO was related to relevant external variables and cognitive measures. Family-level variables, capturing a child's opportunity for social interaction, systematically influenced gaze understanding. Even though the effects were small and confidence intervals were wide, it is remarkable that we were able to detect relationships between this fundamental social-cognitive ability and very distant, real-life variables. In addition, we assessed the influence of gaze understanding on receptive vocabulary. We found a substantial relationship between the two variables, even when correcting for age. Taken together, this speaks to the validity of the TANGO.

Limitations

First, we want to address the scope and interpretation of the TANGO. We believe that solving the task requires locating the attentional focus of an agent as the gaze cues the target location. This speaks to the face validity of the TANGO and its focus on an inherently social stimulus. However, we do not want to claim that the TANGO does not also recruit other, domain-general processes. For example, we believe that a considerable part of gaze understanding relies on vector-following: not just in our task but also in real life. From that perspective, gaze understanding could be seen as a particular case of vector-following that is learned and used in social interactions. Future research could assess how much variation of the gaze understanding task is shared with a physical vector-following task. In addition, computational cognitive models might prove helpful in defining children's behavior on a process-level and disentangling parameters that influence task performance (e.g., spatial acuity).

Second, the influence of testing modality requires further attention. Remote data collection loosens the standardization of the experimental procedure, as we cannot prevent caregivers from interfering. Steering the child's behavior becomes less possible when touchscreens are used, and the child can click on the screen directly. This is why we recommend using tablets for remote data collection. However, it should be noted that families' access to technological devices varies, both across socio-environmental as well as cultural settings.

Third, the children in our sample live in an industrialized, urban Central-European context. It is unclear how our results would generalize to different socio-cultural contexts. A related limitation is that we did not collect demographical information on a participant-level and, instead, had to rely on a community-level description of the sample. This is important to keep in mind when gauging the generalizability of our new measure.

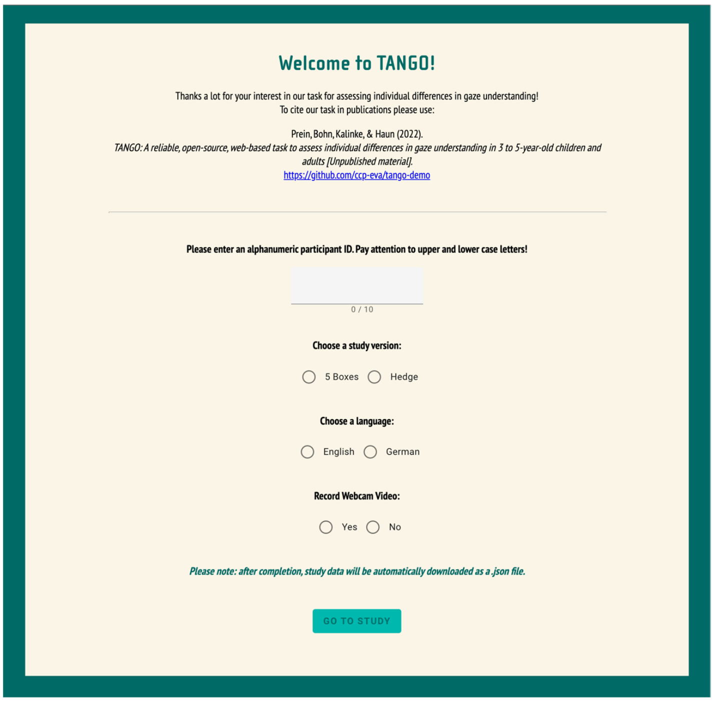

Fig. 5 TANGO demo website. We want to highlight that researchers are welcome to use and modify our task according to their needs. The number of training and test trials and the number of boxes can be adjusted within the JavaScript code, while agents and targets can be exchanged within the HTML code

Finally, we utilized subtle gaze cues in order to increase difficulty and capture individual differences. However, in real-life settings, children could be more accustomed to a combination of head and eye orientation changes, and subtle gaze differences might be less common.

Conclusions

We have presented a new experimental paradigm to study gaze understanding across the lifespan. The TANGO captures individual differences and shows highly satisfactory psychometric properties with respect to internal consistency and test–retest reliability. The browser-based testing infrastructure allows for standardized, portable data collection at scale, both remotely as well as in person. Associations with social-environmental factors and language skills illustrate the validity of the task. Ultimately, this work shows a promising way forward toward more precise measures of cognitive development. The data sets 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 (see Fig. 5): https://ccp-odc.eva.mpg.de/tango-demo/. The code base and respective assets can be accessed in the following repository: https://github.com/ccp-eva/tango-demo. These resources allow interested researchers to use, extend and adapt the task.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02159-5.

Author note The authors made the following contributions. Julia Christin 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 B. M. Haun: Conceptualization, Writing - Review & Editing.

Funding Open Access funding enabled and organized by Projekt DEAL. 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, caregivers, and adults who participated in the study. We thank Jana Jurkat for her help with data collection.

Open practices statement The web application (https://ccp-odc.eva.mpg.de/tango-demo/) described here is open-source (https://github.com/ccp-eva/tango-demo). The data sets generated during and/or analyzed during the current study are available in the [gazecues-methods] repository (https://github.com/ccp-eva/gazecues-methods). All experiments were pre-registered (https://osf.io/zjhsc/).

Declarations

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.

References

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

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

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

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., Bolz, M., Henrich, J., Setoh, P., Wang, J., & Laurence, S. (2013). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society B: Biological Sciences, 280*(1755), 20122654. 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, 10,* 2905. https://doi.org/10.3389/fpsyg.2019.02905

Begeer, S., Bernstein, D. M., van Wijhe, J., Scheeren, A. M., & Koot, H. M. (2012). A continuous false belief task reveals egocentric biases in children and adolescents with autism spectrum disorders. *Autism, 16*(4), 357–366. https://doi.org/10.1177/1362361311434545

Benson, J. E., Sabbagh, M. A., Carlson, S. M., & Zelazo, P. D. (2013). Individual differences in executive functioning predict preschoolers' improvement from theory-of-mind training. *Developmental Psychology, 49*(9), 1615–1627. https://doi.org/10.1037/a0031056

Bernstein, D. M., Thornton, W. L., & Sommerville, J. A. (2011). Theory of mind through the ages: Older and middle-aged adults exhibit more errors than do younger adults on a continuous false belief task. *Experimental Aging Research, 37*(5), 481–502. https://doi.org/10.1080/0361073X.2011.619466

Bohn, M., & Köymen, B. (2018). Common ground and development. *Child Development Perspectives, 12*(2), 104–108. https://doi.org/10.1111/cdep.12269

Bohn, M., Le, K. N., Peloquin, B., Köymen, B., & Frank, M. C. (2021a). Children's interpretation of ambiguous pronouns based on prior discourse. *Developmental Science, 24*(3), e13049. https://doi.org/10.1111/desc.13049

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021b). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour, 5*(8), 1046–1054. https://doi.org/10.1038/s41562-021-01145-1

Bohn, M., Prein, J., Koch, T., Bee, R. M., Delikaya, B., Haun, D., & Gagarina, N. (2023). oREV: An item response theory-based open receptive vocabulary task for 3- to 8-year-old children. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02169-3

Bohn, M., Tessler, M. H., Kordt, C., Hausmann, T., & Frank, M. C. (2023). An individual differences perspective on pragmatic abilities in the preschool years. *Developmental Science,* e13401. Advance online publication. https://doi.org/10.1111/desc.13401

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(*1762). https://doi.org/10.3389/fpsyg.2016.01762

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

Bürkner, P.-C. (2017). brms: an R package for bayesian multilevel models using stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Buttelmann, D., Kühn, K., & Zmyj, N. (2022). The relations among theory of mind, inhibitory control, and aggressive behavior in 4-year-old children – A multi-measure multi-informant approach. *Journal of Cognition and Development, 23*(1), 111–134. https://doi.org/10.1080/15248372.2021.1987240

Byers-Heinlein, K., Tsui, R. K.-Y., van Renswoude, D., Black, A. K., Barr, R., Brown, A., … Singh, L. (2021). The development of gaze following in monolingual and bilingual infants: A multi-laboratory study. *Infancy*, *26*(1), 4–38. https://doi.org/10.1111/infa.12360

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., Tapanya, S., & Singh, S. (2005). Synchrony in the onset of mental-state reasoning: evidence from five cultures. *Psychological Science, 16*(5), 378–384. https://doi.org/10.1111/j.0956-7976.2005.01544.x

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*(4), 1032–1053. https://doi.org/10.1111/1467-8624.00333

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

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. https://doi.org/10.1111/j.1467-8624.2005.00832.x

Coburn, P. I., Bernstein, D. M., & Begeer, S. (2015). A new paper and pencil task reveals adult false belief reasoning bias. *Psychological Research, 79*(5), 739–749. https://doi.org/10.1007/s00426-014-0606-0

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

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.

Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., & Yurovsky, D. (2016). Using Tablets to Collect Data From Young Children. *Journal of Cognition and Development, 17*(1), 1–17. https://doi.org/10.1080/15248372.2015.1061528

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

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The Structure of Working Memory From 4 to 15 Years of Age. *Developmental Psychology, 40*, 177–190. https://doi.org/10.1037/0012-1649.40.2.177

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

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. (2020). *Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/xr7y3

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., & 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., & 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

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., 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

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

Macdonald, R. G., & Tatler, B. W. (2013). Do as eye say: Gaze cueing and language in a real-world social interaction. *Journal of Vision, 13*(4), 6–6. https://doi.org/10.1167/13.4.6

Mahy, C. E. V., Bernstein, D. M., Gerrard, L. D., & Atance, C. M. (2017). Testing the validity of a continuous false belief task in 3- to 7-year-old children. *Journal of Experimental Child Psychology, 160*, 50–66. https://doi.org/10.1016/j.jecp.2017.03.010

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 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

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. (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

Peterson, & WellmanSlaughter, H. M. 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

Poulin-Dubois, D., Goldman, E. J., Meltzer, A., & Psaradellis, E. (2023). Discontinuity from implicit to explicit theory of mind from infancy to preschool age. *Cognitive Development, 65*, 101273. https://doi.org/10.1016/j.cogdev.2022.101273

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: a comprehensive review and systematic assessment. *Psychonomic Bulletin & Review, 29*(1), 44–54. https://doi.org/10.3758/s13423-021-01948-3

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

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology, 1*(4), 223–235. https://doi.org/10.1038/s44159-022-00037-z

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science, 21*(4), e12593. https://doi.org/10.1111/desc.12593

Slaughter, V., & Repacholi, B. (2003). Introduction: individual differences in theory of mind. What are we investigating? *Individual differences in theory of mind: implications for typical and atypical development.* 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

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

Samuel, S., Legg, E. W., Lurz, R., & Clayton, N. S. (2018). Egocentric bias across mental and non-mental representations in the Sandbox Task. *Quarterly Journal of Experimental Psychology, 71*(11), 2395–2410. https://doi.org/10.1177/1747021817742367

Samuel, S., Legg, E. W., Lurz, R., & Clayton, N. S. (2018). The unreliability of egocentric bias across self–other and memory–belief distinctions in the Sandbox Task. *Royal Society Open Science, 5*(11), 181355. https://doi.org/10.1098/rsos.181355

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*(5). https://doi.org/10.3389/fnint.2010.00005

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. (2023). Reply to Poulin-Dubois et al. (2023): Replication problems concerning both implicit and explicit false belief reasoning greatly reduced the chance of finding longitudinal correlations. *Cognitive Development, 65*, 101294. https://doi.org/10.1016/j.cogdev.2022.101294

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

Sommerville, J. A., Bernstein, D. M., & Meltzoff, A. N. (2013). Measuring Beliefs in Centimeters: Private Knowledge Biases Preschoolers' and Adults' Representation of Others' Beliefs. *Child Development, 84*(6), 1846–1854. https://doi.org/10.1111/cdev.12110

Tomasello, M. (2003). *Constructing a language a usage based theory of language acquisition* (8th ed., p. 388). Harvard University Press.

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., & 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

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

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

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

6 Study II

Variation in gaze following across the life span: A process-level perspective

Variation in gaze following across the life span: A process-level perspective

Authors: Julia Christin Prein[1], Luke Maurits[1], Annika Werwach[1,3,4], Daniel B. M. Haun[1,*], Manuel Bohn[1,2,*]

Affiliations: [1] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [2] Institute of Psychology, Leuphana University Lüneburg, Germany. [3] Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany. [4] Max Planck School of Cognition, Leipzig, Germany. [*] shared senior authorship

ORCiD: *Julia Christin Prein* https://orcid.org/0000-0002-3154-6167

Conflicts of interest: The authors declare that they have no conflict of interest.

Data availability statement: The gaze following task (https://ccp-odc.eva.mpg.de/tango-demo/) is open source (https://github.com/ccp-eva/tango-demo). The data sets generated during and/or analyzed during the current study are available in the following repository (https://github.com/ccp-eva/gazecues-modeling). All experiments and analyses were pre-registered prior to data collection (https://osf.io/zjhsc/).

Acknowledgements: We thank Jana Jurkat for her help with data collection and participant recruitment. We would also like to thank Steven Kalinke for his technical programming support. We thank all the children, caregivers, and adults who participated in the study.

Funding: This study was funded by the Max Planck Society for the Advancement of Science, a noncommercial, publicly financed scientific organization (no grant number). Manuel Bohn was supported by a Jacobs Foundation Research Fellowship (grant no. 2022-1484-00).

Ethical statement: The study obtained ethical clearance by the MPG Ethics commission Munich, Germany, falling under an umbrella ethics application (Appl. No. 2021_45). Informed consent was obtained from all individual participants or their legal guardians. The research adhered to the legal requirements of psychological research with children in Germany.

Research highlights

- Gaze following develops beyond infancy. Highest precision levels in localizing attentional foci are reached in young adulthood with a slight decrease towards old age.
- We present a computational model that describes gaze following as a process of estimating pupil angles and the corresponding gaze vectors.
- The model explains individual differences and recovers signature patterns in the data. To estimate the relation between gaze- and vector following, we designed a non-social vector following task.
- We found substantial correlations between gaze following and vector following, as well as Level 2 perspective-taking. Other Theory of Mind tasks did not correlate.

Abstract

Following eye gaze is fundamental for many social-cognitive abilities, for example, when judging what another agent can or cannot know. While the emergence of gaze following has been thoroughly studied on a group level, we know little about (a) the developmental trajectory beyond infancy and (b) the sources of individual differences. In Study 1, we examined gaze following across the lifespan (*N* = 471 3- to 80-year-olds; children from a mid-sized German city; international remotely tested adults). We found a steep performance improvement during preschool years, in which children became more precise in locating the attentional focus of an agent. Precision levels then stayed comparably stable throughout adulthood with a minor decline toward old age. In Study 2, we formalized the process of gaze following in a computational cognitive model that allowed us to conceptualize individual differences in a psychologically meaningful way (*N* = 60 3- to 5-year-olds, 50 adults). According to our model, participants estimate pupil angles with varying levels of precision based on observing the pupil location within the agent's eyes. In Study 3, we empirically tested how gaze following relates to vector following in non-social settings and perspective-taking abilities (*N* = 102 4- to 5-year-olds). We found that gaze following is associated with both of these abilities but less so with other Theory of Mind tasks. This work illustrates how the combination of reliable measurement instruments and formal theoretical models allows us to explore the in(ter)dependence of core social-cognitive processes in greater detail.

Keywords: social-cognitive development, theory of mind, gaze following, individual differences, cognitive modeling, lifespan

Word count: 10469

Introduction

Following the gaze of others is valuable for extracting information from the environment. It guides us to “informational hotspots” (Meltzoff et al., 2010, p. 1) and can be used to identify internal states such as intentions or emotions (Corkum & Moore, 1998; Pfeiffer et al., 2013). As one of the most fundamental social-cognitive abilities, gaze following is an integral part of almost every form of social interaction, including communication, collaboration, cultural and social learning (Bohn & Frank, 2019; Emery, 2000; Frith & Frith, 2012; Hessels, 2020; Moore, 2008; Rakoczy, 2022; Tomasello & Rakoczy, 2003), and has been extensively studied in infancy (for review, see Del Bianco et al., 2019). In the most commonly used paradigm (e.g., Astor et al., 2020; Byers-Heinlein et al., 2021; Gredebäck et al., 2010; Ishikawa et al., 2022), the experimenter looks directly at the infant before shifting their head and eyes to one of two objects. Infants’ looking times to the target or the proportion of choosing the target over the distractor are measured. Research in this tradition finds that infants as young as three to four months can follow the gaze of another agent (Astor et al., 2021; D’Entremont et al., 1997; D’Entremont, 2000; Del Bianco et al., 2019).

Previous research has shown a refinement of gaze following abilities in a child’s first and second year of life (e.g., Astor et al., 2021; Brooks & Meltzoff, 2002; Butterworth & Jarrett, 1991). At the end of their first year of life, infants can follow gaze to locations outside their current visual field and move themselves to gain proper perceptual access (Butterworth & Jarrett, 1991; Corkum & Moore, 1995; Deák et al., 2000; Moll & Tomasello, 2004). However, we do not know much about the developmental progression beyond these qualitative milestones. One possibility is that the ability to follow gaze does not improve beyond infancy. Yet, most, if not all, cognitive abilities continue to develop throughout childhood (e.g., Gathercole et al., 2004;

Gredebäck et al., 2010). It seems likely that children fine-tune their gaze following as they get older - presumably while using them in social interactions. To capture the development in gaze following beyond infancy, Study 1 included participants from preschool to old age.

Up until today, only a handful of gaze following studies differentiate between manipulating head and eye movement. Michel et al. (2021) found that gaze following in four-month-olds was likely driven by other's head- instead of eye movements. Corkum and Moore (1995), Lempers (1979), and Lempers et al. (1977) suggest that infants, at least until 19 months, struggle when eye and head direction diverge. From farther distance, body or face orientation can act as more salient cues to determine another's area of attention. However, eye direction indicates a more precise location of focus (Emery, 2000; Stiefelhagen & Zhu, 2002; however, see Loomis et al., 2008 for peripheral vision), and allows to anticipate likely future actions (Friesen & Rao, 2011; Zohary et al., 2022). The three studies included in this work, therefore, focused on subtle gaze cues and isolated eye movement alone.

Group-level analyses of gaze following abilities (e.g., average age at which children as a group reach an above-chance performance) may mask individual differences between children. Measuring individual differences in basic aspects of social cognition is important to understand the underlying processes and to quantify the impact of environmental influences and other cognitive abilities (Birch et al., 2017; Del Bianco et al., 2019). Across the three studies, we measured gaze following continuously by using a task that is designed to capture individual-level variation (Prein et al., 2023): the TANGO (Task for Assessing iNdividual differences in Gaze understanding - Open) avoids floor and ceiling effects in children and adults and is thus particularly suited to examine how gaze following changes with age.

A promising approach to interpreting individual differences in cognitive abilities is computational cognitive modeling. Existing computational models have described gaze following via reinforcement learning (Ishikawa et al., 2020), as a consequence of goal interference and mapping self-experience onto other agents (Friesen & Rao, 2011) or an interplay of object distances/saliencies and head poses (Jasso & Triesch, 2006; Lau & Triesch, 2004; Recasens et al., 2015). As such, these models focus on the motivation behind gaze following or on situations in which the context (e.g., head orientation, surrounding objects) offers cues to the gaze direction. Our goal, however, was to formulate a theory that models how people estimate gaze direction based on the eyes alone (i.e., "literal" gaze following) when no target objects are present [1] and the other's head is frontally oriented.

To our knowledge, there are three views that focus on the eyes alone that conceptualize gaze as (1) a beam, (2) a cone, or (3) a line (alternatively called a vector, ray, or line-of-sight). Guterstam et al. (2019) proposed the idea of gaze as a force-carrying beam. This theory, however, focuses on how people's implicit assumptions about the physical properties of an object changes when someone looks at it. The idea of gaze as a cone is mostly concerned with the question of how people determine whether someone else looks at them (Gamer & Hecht, 2007; Horstmann & Linke, 2021). The conception most relevant to the present study sees eye gaze as a line: Yaniv and Shatz (1990) proposed that children extrapolate an imaginary trajectory between the agent

[1] As Jasso and Triesch (2006) put it: "There is a general agreement that tests of gaze following in the absence of targets are more stringent than with targets, because that eliminates the possibility that infants are simply following the target's saliency" (p. 2).

and the object to identify the focus of attention (similar to "geometrical" gaze following; (Butterworth & Jarrett, 1991)). Anecdotal evidence was already reported by Walker and Gollin (1977), who observed two children pointing their fingers into the air, drawing a line between an agent and an object, and saying, "He sees that" (p. 354). Michelon and Zacks (2006) found that response time in Level 1 visual perspective-taking tasks depends on the distance between the agent and the object (i.e., the length of the line-of-sight). Some researchers additionally highlight iris eccentricity, that is, the ratio of the visible sclera on each side of the pupil (S. M. Anstis et al., 1969; Symons et al., 2004; Todorović, 2006). Todorović (2006) defined gaze direction as "the vector positioned along the visual axis, pointing from the fovea of the looker through the center of the pupil to the gazed-at spot" (p. 3550). Symons et al. (2004) furthermore reasoned that "the perceiver must use the asymmetrical configuration of the dark-white contrast of another individual's eyes, and trace along two invisible sight-lines to their convergent point, that is, the third part of the triad (e.g., an object or a person)" (p. 452). Based on their finding that adults' gaze direction sensitivity decreases when only one eye is shown, Symons et al. (2004) conclude that information from both eyes must be integrated.

While the previously mentioned conceptualizations focus on the direction of eye gaze, they (A) cannot explain how people differ in their abilities to precisely estimate gaze direction, and (B) are not clearly expressed as formal, mathematical models with explicit assumptions and testable predictions. In the words of (Gamer & Hecht, 2007): "Given the social relevance of determining gaze direction, the psychophysics of gaze is underdeveloped" (p. 705).

Here, we propose a cognitive model of gaze following, which builds upon the notion of eye gaze as line-of-sight tracing and extends this by explicitly modeling individual differences. Our gaze model assumes participants infer the locus of someone's attention to be where two estimated gaze vectors meet. Each of these gaze vectors results from connecting the center of the agent's eyeball and the center of the pupil. Because the center of the eyeball is not directly observable, vector estimation happens with a degree of uncertainty. Development of gaze following corresponds to a decrease in uncertainty. Individual differences correspond to systematic differences in uncertainty.

By focusing on individual differences, we can further address the relationship between gaze following and other cognitive abilities. A longstanding question has been whether gaze following is related to Theory of Mind (ToM) (e.g., Brooks & Meltzoff, 2015). Moll and Meltzoff (2011) have suggested that joint attention (including gaze following) might be seen as "Level 0 perspective-taking", which provides the foundation for later-emerging, more complex perspective-taking abilities. On the other hand, an alignment of infants' visual attention to another's gaze does not necessarily indicate understanding the intentions of the agent (Aslin, 2007). Infants could simply align their orientation without processing what exactly the other is seeing (Butterworth & Jarrett, 1991). In fact, one might question if such an alignment reflects an understanding of visual perspectives at all because the "target" or "object of representation" is not necessarily specified (Perner et al., 2003, p. 358). Consequently, Astor and Gredebäck (2022) have listed the relationship between gaze following and perspective-taking as one of their five big open questions in gaze following research. Therefore, in Study 3, we assessed how gaze following relates to ToM abilities, especially visual perspective-taking.

Taken together, the present study had three main goals: first, we studied the development of gaze following beyond infancy (Study 1). Instead of capturing the youngest age at which children follow gaze, we examined how this ability changes with age. Our second goal was to provide a process-level theory of gaze following – and, most importantly, individual differences therein. We proposed a computational cognitive model, which formalized gaze following as a form of vector following, and tested whether our model explained empirical data (Study 2). Third, we examined which (social-)cognitive components comprise gaze following (Study 3). Based on our model, we predicted that gaze following should be related to non-social vector following. Additionally, we assessed the link between gaze following and ToM measures, with a particular focus on visual perspective-taking.

Study 1: Gaze following across the lifespan

The study was pre-registered prior to data collection: https://osf.io/snju6 (child sample) and https://osf.io/6yjz3 (adult sample). The study obtained ethical clearance by the MPG Ethics commission in Munich, Germany, falling under an umbrella ethics application (Appl. No. 2021_45). Data was collected between May 2021 and April 2023.

Participants

We collected data online from 3- to 80- year-olds (see Supplements for further details). The child sample consisted of 471 participants and was recruited via an internal database of families in Leipzig, Germany, who volunteered to participate in child development studies. Participants came from ethnically homogeneous, mixed socioeconomic backgrounds with mid to high parental education levels. They lived in an industrialized, urban Central-European context in a mid-size German city (approx. 600,000 inhabitants; median individual monthly net income approx. 1,600€ as of 2021). Most were raised monolingually in a nuclear two-generational family setting. Information on demographics and socioeconomic status was not formally recorded on a participant level.

Adults were recruited via *Prolific* (Palan & Schitter, 2018). *Prolific* is an online participant recruitment tool from the University of Oxford with predominantly European and US-American subjects. Participants consisted of 240 English-speaking adults who reported to have normal or corrected-to-normal vision. For completing the study, subjects were paid above the fixed minimum wage (~£10.00/hour).

Materials

We used the continuous version of the TANGO (Prein et al., 2023). The task was presented as a web application (demo https://ccp-odc.eva.mpg.de/tango-demo/; source code https://github.com/ccp-eva/tango-demo). The TANGO showed satisfactory internal consistency and retest reliability (*Pearson's r* from .7 to .8; Prein et al. (2023)) and no floor or ceiling effects for children and adults.

Procedure

Children and teenagers received a personalized link to the study website. Caregivers were asked to provide technical support, while explicitly being reminded not to help in responding. Webcam videos were recorded whenever consented and technically feasible in order to monitor whether children and teenagers responded on their own. Adults completed the online study unsupervised.

Each trial presented an agent standing in a window, watching a balloon (*i.e.*, target) falling to the ground (see Figure 4A; however, Study 1 presented animal agents). The target fell behind a hedge while the agent's gaze followed the target's trajectory. In test trials, a hedge covered the target's position. Participants were asked to touch or click where they estimated the target to be based on the agent's gaze. Four familiarization trials ensured participants understood the task and felt comfortable with the response format. Then, 15 test trials followed. Completing 19 trials took 5-10 minutes.

We measured imprecision, defined as the absolute difference between the target center and the x coordinate of the participant's click. The screen width was divided into ten bins. Within each bin, exact target coordinates were randomly generated. Each target bin, agent, and target color occurred equally often and did not appear in more than two consecutive trials.

Analysis

We ran all analyses in R version 4.3.3 (2024-02-29) (R Core Team, 2022). Regression models were fit as Bayesian generalized linear mixed models (GLMMs) with default priors using the function `brm` from the package `brms` (Bürkner, 2017, 2018).

We fit GLMMs that make different assumptions about the developmental trajectory, modeling the relationship between age and performance as linear, quadratic, or cubic. In addition, we fit a Gaussian Process model (Bürkner, 2017), which assumes a smooth relationship but avoids enforcing a particular shape. Per individual, imprecision was aggregated across trials and modeled as a `lognormal` distribution.[2] The unit of imprecision was counted in target widths, i.e.,

[2] Originally, we fit our models on a trial-by-trial basis with the following structure: `performance ~ age + symmetricPosition + trialNr + (1 + symmetricPosition + trialNr | subjID)`. However, the Gaussian Process model was computationally heavy. Thus, we simplified the model structure, aggregated data on a subject level, and included only age as an effect. See

an imprecision of 1 meant clicking one balloon width to the left or right of the true target center. We inspected the posterior distributions (mean and 95% Credible Interval (CrI)) for the age estimates and compared models via model weights and the difference in expected log pointwise predictive density (ELPD) estimated using leave-one-out cross-validation (LOO) (Vehtari et al., 2017).

To obtain a concise but principled characterization of the developmental trajectory, we additionally performed a Bayesian change point analysis, using the package `RBeast` (Zhao et al., 2019). We sought the most likely change points in our data, assuming a constant mean (i.e., a flat line, zero-degree polynomial) within each segment. To avoid "overreactions" to outlying data points, we constrained the model to have minimally 10 data points between consecutive change points (= half of the data points collected per adult decade). We inspected the posterior probability of different numbers of change points and the locations of these change points (mean and 95% CrI).[3]

Supplements for a comparison between the original and the here-reported model structures. The model predictions did not differ notably.

[3] In a supplementary analysis, we varied the parameters of our change point analysis and modified the number of allowed change points, the minimum number of data points between change points, and the polynomial order. When we allowed more explorative room, the models became more sensitive and added more fine-grained change points. The exact location of the change points varied slightly. However, the overall interpretation stayed the same, fitting our initial visual inspection: while early childhood was characterized by much change, adults showed

Result

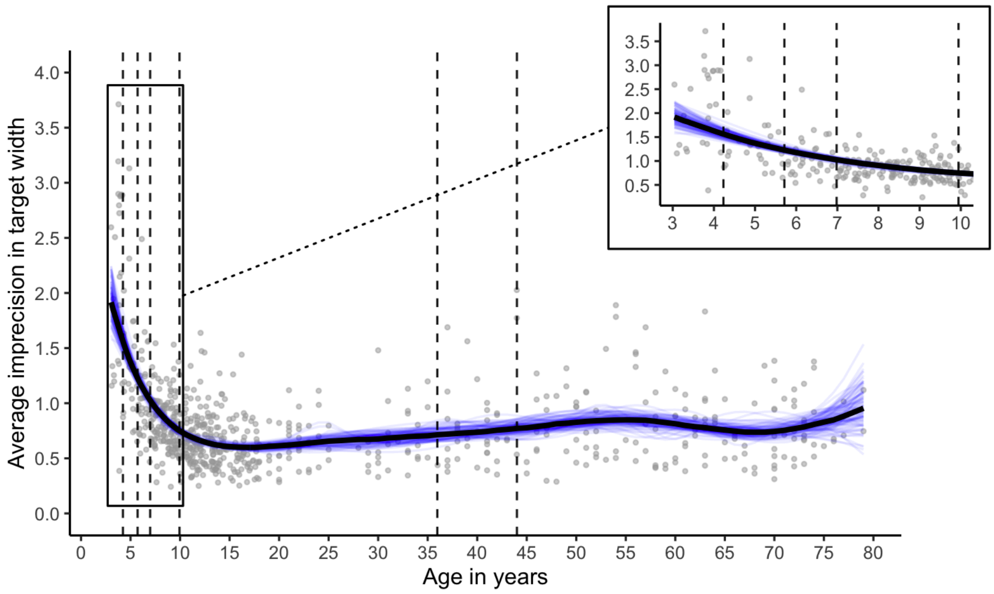

Figure 1: **Developmental trajectory of gaze following across the lifespan**. Grey dots show the mean level of imprecision (ie., absolute distance between the target's center and the participant's click) for each subject averaged across trials. The unit of imprecision is counted in target width, i.e., a participant with imprecision of 1 clicked on average one target width to the left or right of the true target center. Blue lines show 100 draws from the expectation of the posterior predictive distribution of the Gaussian Process model, with its mean predicted developmental trajectory as a

a relatively stable level of imprecision, with a minor decay toward elderly adulthood. See Supplements for further detail.

solid black line. Vertical, black, dashed lines show the locations of the changes with the highest posterior probability according to our Bayesian change point analysis.

High levels of variation pointed to substantial individual differences in all age groups (overall imprecision mean = 0.81, sd = 0.82, range = [0 - 10.73]). We found substantial evidence for a non-linear development in gaze following across the lifespan. The Gaussian process model was clearly preferred over the polynomial models due to the highest predictive accuracy according to the LOO ELPD estimates: elpd_diff between Gaussian process and cubic model = -33.75 (SE = 8.83); elpd_diff between Gaussian process and quadratic model = -95.07 (SE = 15.19); elpd_diff between Gaussian process and linear model = -127.17 (SE = 18.18); all in favor of the Gaussian process model. Moreover, the Gaussian Process model showed the greatest model weight (approximating 1). For the imprecision in gaze following, the standard deviation of the Gaussian process (SD = 1.52, 95% CrI [0.25; 5.19]) indicated nonlinearity.

The Bayesian change point analysis revealed 6 major shifts in gaze following during the lifespan (MAP estimate with 23.49% probability). The change points occurred at 4.23 years (95% CrI [4.13; 4.33], mean imprecision until change point = 2.15, SD = 0.85); 5.71 years (95% CrI [5.38; 5.90], mean = 1.36, SD = 0.53); 6.98 years (95% CrI [6.64; 8.04], mean = 1.06, SD = 0.40); 9.94 years (95% CrI [8.66; 12.55], mean = 0.82, SD = 0.23); 35.97 years (95% CrI [27.30; 39.69], mean = 0.65, SD = 0.23); and 44.01 years (95% CrI [40.37; 44.43], mean = 0.79, SD = 0.40). In short: we found a rapid initial improvement in gaze following in early childhood, followed by a long period of minor, very slow change with slightly increasing levels of imprecision toward old age.

Discussion

We investigated the shape of change in gaze following across the lifespan and found a non-linear developmental trajectory, in which young children quickly enhanced their level of proficiency. Performance peaked around early adulthood, while there was a minor decay in later adulthood. These results support the idea that humans fine-tune their existing gaze following ability after the first emergence in infancy. Furthermore, we observed substantial individual differences in all age groups. While variation was highest in the three- and four-year-olds, it remained relatively stable across the lifespan.

Previous studies found that already four-month-olds demonstrate basic gaze following abilities (Del Bianco et al., 2019). Since we measured an active location choice on a touchscreen, we could not collect data from infants. In our sample, three-year-olds were still rather imprecise in their gaze following ability (average imprecision was approx. two target widths). How can we explain this divergence? First, we used subtle eye movements as cues. Many existing studies let the agents move eye and head in parallel (Behne et al., 2005; Povinelli et al., 1997), establishing a confound with the more salient head movement. Relying exclusively on eye movements might be more difficult for children than presenting them with a combined eye and head orientation (Carpenter et al., 1998). Silverstein et al. (2021) used a similar manipulation of gaze cues without head rotation and found that 6- to 18-month-olds were around or just above chance for gaze following. The authors argue that infants might fixate on another's face most of the time, while eye movement alone might not be strong enough to guide their attention. Furthermore, our study required participants to (1) precisely follow an agent's gaze, (2) interpret this as a cue, and (3) use the cue to guide their own behavior. It is conceivable that three-year-olds followed the agent's gaze but did not translate this into precise, active behavior. Moll and Kadipasaoglu (2013) argue that social forms of perspective-taking evolve prior to visual perspective-taking, which only emerges within the third year of life. Young children might simply not be interested in a differential, spatial representation of the surrounding objects. Taken together, this might explain why our younger participants located the agent's gaze rather imprecisely.

Regarding our sample of elderly adults, we expect a sampling bias (Bethlehem, 2010; Gosling et al., 2004; Remillard et al., 2014). First, certainly not all older people have a high-speed internet connection or are knowledgeable in its use. Second, the elderly adults participating in *Prolific* studies might show greater cognitive flexibility compared to their offline counterparts. Therefore, a representative sample may show a greater age decline in gaze following compared to our reported sample. In addition, older people might be more likely to suffer from visual impairments. Even though we filtered participants to only include normal- to correct-to-normal vision, we cannot guarantee that our participants showed no symptoms of reduced vision.

Study 2: Computational cognitive model

Our lifespan study showed that gaze following develops throughout childhood, and variation between individuals appears in all age groups. The TANGO has previously been shown to reliably capture inter-individual differences in gaze following (Prein et al., 2023). The variation between participants was thus likely genuine and not due to random noise. In Study 2, we aimed to understand the developmental change and individual differences on a process level. We present a theory of gaze following that explains how participants process the available gaze information and trace a line-of-sight to identify the agent's focus. We formalized this inference process in a computational cognitive model that replicates a schematic representation of how participants make inferences in the task's context (i.e., a model of the task and not the data).

The study design and procedure obtained ethical clearance in the same way as Study 1. The study and the model were pre-registered prior to data collection: https://osf.io/r3bhn. Data were collected between May and August 2021.

Computational model

Our model quantifies a participant's cognitive ability to follow gaze by inverting a probabilistic process that generates the participant's clicks from observing the eyes of the agent. It is formally defined as:

$$P(\theta|x_c, \alpha_l, \alpha_r) \propto P(x_c|\alpha_l, \alpha_r, \theta)P(\theta)$$

where θ is an individual's cognitive ability to locate the focus of the agent's attention, x_c is the coordinate the participant clicked, and α_l and α_r are the pupil angles for the left and right eye, respectively. The pupil angle α is defined as the angle between a line connecting the center of the eye to the pupil and a line extended vertically downward from the center of the eye (see Figure 2A)[4]. Please note that in our case, the center of the pupil is simultaneously the center of the iris (see S. Anstis (2018) for the influence of moving irises, pupils, and corneal reflexes).

---

[4] This model mirrors the logic of the TANGO programming code. In the online experiment, we read out the center point coordinates of the target and the agent's eyeball (i.e., the SVG coordinates), and then calculated a line between these two points: this was our gaze vector (acting in the functionally same way as a pupil angle). Knowing the eyeball radius, we calculated the point of intersection at which the gaze vector met the eyeball boundary. Finally, the agent's pupil moved from the center of the eyeball along the gaze vector to the intersection point. This way, the

Based on our verbal task instructions, we assumed that participants (1) expected the agent's looks to be directed at the target, and (2) to click on the coordinate they estimated the agent to look at. Consequently, we did not assume that participants' clicks were noisy in any way but that they clicked on the screen location where they genuinely thought the target was (and that the agent was looking at).

The true eye angles (α_l and α_r) cannot be directly observed and have to be estimated based on the position of the pupils within the eyes, resulting in approximate values ($\widehat{\alpha_l}$ and $\widehat{\alpha_r}$). We presumed this estimation to be a noisy process. Thus, we conceptualized the development of the cognitive ability to follow gaze as a reduction of noise in the estimates (i.e., an increased certainty about the pupil angles).

Any clicked value of x_c implied a "matched pair" of the estimated pupil angles $\widehat{\alpha_l}$ and $\widehat{\alpha_r}$, with the property that lines extended along those two angles met at the precise location of the target. As a consequence, we can rewrite the likelihood function of the model above:

$$P(x_c|\alpha_l, \alpha_r, \theta) \propto P(\widehat{\alpha_l}, \widehat{\alpha_r}|\alpha_l, \alpha_r, \theta)P(x_c)$$

$P(x_c)$ is a prior over potential target locations, which we assumed to be skewed towards the screen center: We anticipated that participants have an a priori expectation that the target will land close to the middle, because the target was last visible in the screen center before disappearing behind the hedge and because the agent was located centrally on the screen. We

agent was animated to "look at" the target. In the gaze model, we assumed participants go through these steps in reverse order.

estimated the strength of this center bias (i.e., the standard deviation of a Normal distribution around the screen center) based on the data: $P(x_c) \sim \mathcal{N}(960, \sigma^p)$.

The width of this distribution is defined by σ^p. For children, we assumed that the center bias changed with age and estimated σ^p via a linear regression as a function of the child's age (age_i): $\sigma^p = \beta_0^{\sigma^p} + age_i \cdot \beta_1^{\sigma^p}$. Therefore, the participant-specific distribution for $P(x_c)$ was constrained by the performance in the TANGO and the child's age. For the adults, σ^p was not age-specific.

The main inferential task for the participant lay in estimating the pupil angles, i.e., sampling from the first term of the right-hand side equation above, $P(\widehat{\alpha_l}, \widehat{\alpha_r} | \alpha_l, \alpha_r, \theta)$. For this, we assumed that the pair of estimated pupil angles were sampled from a probability distribution which is the product of two Normal distributions of equal variance, σ_v, centered on the true pupil angles:

$$P(\widehat{\alpha_l}, \widehat{\alpha_r} | \alpha_l, \alpha_r, \theta) \propto \phi(\hat{\alpha}_l; \alpha_l, \sigma_v)\phi(\hat{\alpha}_r; \alpha_r, \sigma_v),$$

As σ_v determined the level of accuracy with which participants estimated the pupil angles, it is the component of the model that defines θ. When σ_v is very small (i.e., the distribution around the pupil angle is narrow), clicks far away from the target are unlikely, as these would require estimated pupil angles very different from the true pupil angles. When σ_v is very large (i.e., the distribution around the pupil angle is wide), almost any pupil angles may be sampled, corresponding to a roughly uniform distribution over click coordinates. We expected σ_v to vary between individuals. Consequently, individuals differed in the level of precision with which they can locate the target based on observing the agent's eyes.

The shape of the $P(x_c|\alpha_l, \alpha_r, \theta)$ distribution leads to a testable group-level prediction. As the pupil location varies, a fixed amount of uncertainty around the pupil angle corresponds to a varying degree of uncertainty in the estimated target location (see Figure 2B & C). When the agent directs their gaze toward the very left or right side, the distribution around the target location from which participants sample is comparatively wider than when the agent gazes centrally to the ground. For illustrative purposes, imagine a similar phenomenon: pointing a torch light to a flat surface on the ground. When one points the light cone directly at the surface, the light beam is concentrated in a clearly defined, small, symmetric area. When one points the light cone further away from oneself (shining at an angle), the light from one half of the cone must travel further to reach the surface than the light from the other half, resulting in an asymmetric light pattern. As the angle increases, the light is spread over a wider area, and the surface is illuminated less evenly. Consequently, for the same σ_v, the further out a target coordinate lies, the wider and less symmetric the distribution. This increases both the variance and the bias in a participant's estimate of the agent's attentional focus, resulting in a decreased performance in the task. As σ_v decreases and the cone narrows, the extent to which performance varies at different angles decreases.

Our gaze model consequently predicted that TANGO trials vary in difficulty (see Figure 2B and C): participants should be more imprecise in locating the target the further out it lands, resulting in a U-shaped pattern[5]. If our data matched the pattern of this model prediction, this

---

[5] In our screen-based study, this effect should decrease again towards the most outward sides. Since the computer screen has a natural border, trials in which the target lands furthest out to the

could act as evidence for the gaze model. Therefore, our gaze model provides a quantitative theory of gaze following. In the following, we tested these predictions in children and adults.

left/right become slightly easier again. In these cases, the uncertainty about the pupil angle faces practically only the inner side (facing the center) of the screen, since the natural border of the screen limits where participants can click. In another adult sample with more trials, we could recover this pattern. For further elaboration, see Supplements.

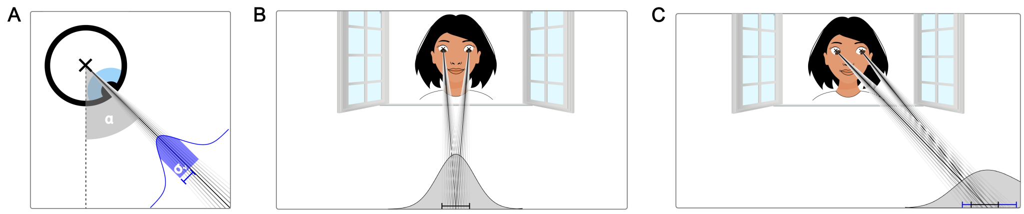

Figure 2: **Gaze model** (A) Visualization of the gaze model (simplified, for one eye). Participants are assumed to observe the pupil location and estimate the center of the agent's eye. Connecting these two point estimates as a line yields the unique vector that extends from the center of the agent's eyeball through the center of the pupil to the attentional focal point (line-of-sight). The angle between this vector and a line pointing vertically to the ground (black dashed line) is the pupil angle (α). Participants are assumed to sample (grey lines) from Normal distributions (blue line) centered around the true pupil angle. The variance of the Normal distribution (σ_v) is expected to vary between participants. (B & C) Geometrical features of the gaze model. As the pupil angle varies, a fixed amount of uncertainty in the angle corresponds to a varying degree of uncertainty in the estimated target location. The distribution around the target location from which participants sample is wider when the agent gazes toward the side than when she gazes centrally. The blue line on the ground shows the added level of uncertainty in the estimated target position for the target location further outward (C).

Participants

The sample consisted of 60 children, including 20 three-year-olds (mean age = 3.47 years, SD = 0.34, range = 3.07 - 3.97, 11 girls), 20 four-year-olds (mean age = 4.61 years, SD = 0.26, range = 4.09 - 4.98, 10 girls), 20 five-year-olds (mean age = 5.66 years, SD = 0.24, range = 5.01 - 5.96, 12 girls). Children were recruited via an internal database, where each parent previously consented to child development studies, and data was collected in kindergartens in Leipzig, Germany.

In addition, we included 50 adults from Study 1 (mean age = 31.92 years, SD = 12.15, range = 18 - 63, 36 female). Since developmental change was minimal in our adult sample (see Study 1) and the cognitive models were computationally heavy, we decided to only include the first 50 adults who had completed the study.

Procedure

We applied the same procedure as in Study 1. Children were tested in a quiet room in their kindergarten, while an experimenter guided the child through the study on a tablet. Adults participated online.

Analysis

We quantified how well our gaze model explained the gaze following process in two ways. First, we aggregated the model predictions and data for each target bin and age group (3-, 4-, 5-years-olds, adults), and computed correlation to quantify how well the model was able to recover the data. Second, we compared the predictions of our gaze model to two simple alternative models that assume participants do not rely on the agent's gaze at all: a random guessing model and a center bias model. The random guessing model assumed participants randomly clicked on the screen and was implemented as sampling from a uniform distribution over all possible coordinates, $\mathcal{U}(0,1920)$. The center bias model assumed participants always clicked near the screen center and was implemented as sampling from a Normal distribution with the screen center as the mean, and one balloon width as the variance, $\mathcal{N}(960,160)$. Note that the center bias model also predicted imprecision should be higher for targets further out on the screen. However, compared to the gaze model, it predicted a steep effect towards the sides, resulting in a V-shaped pattern (imprecision as the distance between the target location and the screen center). All cognitive models were implemented in WebPPL (Goodman & Stuhlmüller, 2014).[6]

We compared models via the marginal likelihood of the data under each model. The pairwise ratio of marginal likelihoods for two models is also known as the Bayes Factor, which quantifies the quality of a model's predictions by averaging over the possible values of the model's parameters weighted by the prior probabilities of those parameter values. It can be used to estimate how much more likely the data under one model are compared to the other. Bayes Factors implicitly consider model complexity: models with more parameters often have broader prior distributions over parameters, which might weaken potential gains in predictive accuracy.

Results

We found very clear support for our gaze model, both in children as well as adults. A strong correlation between the data mean and the gaze model estimate (σ_v) showed that the data mean is suitable to quantify individual differences: child sample r = 0.95, 95%CI [0.92, 0.97];

[6] In an exploratory analysis, we simulated data for two more alternatives: a line-of-sight tracing model that assumed no inferential noise in the participants' gaze following ability, and a model building up on the line-of-sight tracing with added motor noise. Please note that these did not predict a U-shape pattern, since all target locations would be influenced by the motor noise equally. For further details, see Supplements.

adult sample r = 0.96, 95%CI [0.93, 0.98]). The gaze model predicted a U-shaped pattern which we also observed in our data (see Figure 3C). The model comparison strongly favored our gaze model over the center bias model (child sample $logBF_{10}$ = 1,015.33; adult sample $logBF_{10}$ = 2,575.75) and the random guessing model (child sample $logBF_{10}$ = 388.98; adult sample $logBF_{10}$ = 919.03). For the child sample, we went one step further into the analysis. When correlating the observed data across all target positions with the predictions of the three models, we found a high similarity for the gaze model: r = 0.95, 95%CI [0.90, 0.98], while the correlations with the alternative models were smaller (center bias model: r = 0.77, 95%CI [0.57, 0.89]; random guessing model: r = 0.78, 95%CI [0.58, 0.89]). The gaze model's prior over potential target locations assumed that participants' clicks would be skewed towards the screen center. For children, we estimated the width of the prior's distribution based on the participants' age. Interestingly, we found that the differential age effects on the center bias prior were minor (intercept = 300.14, 95%HDI [277.19; 321.28]; slope = 1.18, 95%HDI [0.00; 18.78]). The slope of the prior indicated that older children were only marginally less drawn towards the screen center compared to the younger children.

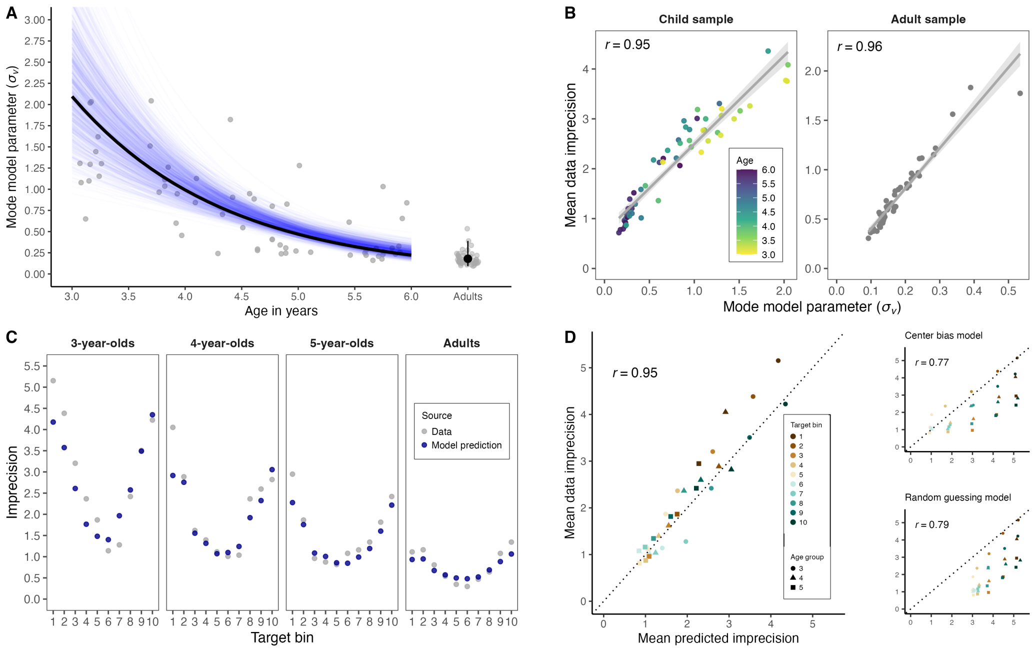

Figure 3: **Gaze model** (A) Developmental trajectory of the estimated model parameter. Grey dots show individual level parameter values. The black line shows the maximum a posteriori (MAP) estimate; blue lines show 1000 draws from it. The large black point with 95% HDI shows the mean of the model parameter for the adult sample. We added minimal horizontal and vertical noise to the adult individual level parameter values to avoid overplotting. (B) Correlation between estimated mode of the model parameter and data mean per individual, faceted by age group. In the child sample, color denotes age in years. Grey regression lines with 95% CI show smooth conditional means based on linear models, with *Pearson*'s correlation coefficient *r*. (C) Pattern recovery. Imprecision in target width for each target bin by age group. Model predictions in blue; data in grey. (D) Correlation between the observed data and the predictions of the three models by target position and age group (across individuals; only for the child sample).

Discussion

We presented a formal cognitive model of gaze following to describe how gaze following develops with age and varies between individuals. We modeled gaze following as a process in which participants estimate pupil angles based on the pupil location within the eye. By following the resulting gaze vector, they consequently arrive at the attentional focus of the agent. Individual differences can be explained as varying levels of imprecision in the pupil angle estimation. We assume the basic process of gaze following to be the same across the lifespan, though individuals become increasingly precise with age. While we did not include an infancy sample, we believe the process of gaze following should operate similarly — if not other cues such as the object saliencies or head directions are followed instead. By conducting model comparisons, we ruled out simpler explanations of the data.

In addition, we observed differences in performance depending on gaze location: participant data showed that precision levels dropped as the agent's gaze moved further away from the center. Our gaze model predictions recovered this "signature pattern" in the data. Future research could use this signature in the data as evidence of whether diverse communities employ the same inferential mechanism to solve the task, speaking for a shared cognitive architecture.

Interestingly, the U-shaped pattern in the TANGO task can be conceptually compared to the result patterns of Michelon and Zacks (2006): in their Level 1 perspective-taking task, an increased distance between the agent and the target decreased performance in adults (i.e., reaction times). Targets closer to the midline were more easily traced than ones further away from the agent. The authors concluded that visual acuity is generally higher for locations on the vertical than the diagonal axis ("oblique effect"; (Appelle, 1972; Heeley et al., 1997; Mikellidou et al., 2015)). Similarly, Symons et al. (2004) have found that adults' acuity for gaze direction depends on the target location. Not only is the increased imprecision for TANGO trials in which the target lands further out consistent with this finding, but our gaze model poses a viable explanation for this effect.

A limitation of our model is that we cannot disentangle how much of the participants' uncertainty comes from a noisy estimate of the agent's attentional focus and how much is due to imprecise clicking (e.g., experiencing motor issues, adding random noise to the click). However, we believe imprecise clicking to be of minor concern, since the children in our sample seemed determined in where they clicked and to not have issues with aiming or motor control (see precision in training trials in Supplements).

A critical feature of our model is that it assumes gaze following to rely on vector following: subjects are modeled to calculate pupil angles which serve as gaze vectors to point to the attentional focus point of an agent. Even though this vector following component is a geometrical calculation, one must first interpret the agent's eyes as a relevant social stimulus. Therefore, our model describes gaze following as a particular form of vector following in a social context.

Study 3: Components of gaze following

Study 3 examined the components of gaze following and whether it can be fully reduced to physical vector following. The positive link between TANGO and social-environmental factors like age of childcare entry (Prein et al., 2023) underline how social interaction is integral to gaze following. Nevertheless, it is unclear how gaze following relates to other forms of perspective-taking (Astor & Gredebäck, 2022). Therefore, we investigated associations between gaze following and other measures of social-cognitive abilities.

First, we experimentally isolated the vector following component of the TANGO. We designed a new non-social vector following task that shared all crucial design features of the TANGO. Second, we assessed children's social-cognitive abilities by administering a ToM task battery, comprising four tasks from the ToM scale by Wellman and Liu (Wellman & Liu, 2004) and two additional perspective-taking tasks (Flavell, Flavell, et al., 1981; Flavell, Everett, et al., 1981). We reasoned that the TANGO shares task demands with the non-social vector following task while it shares its social context with the ToM tasks. We could, therefore, imagine both an absence or presence of relationship between gaze following and ToM. As stated in our pre-registration, we further assessed whether the two perspective-taking tasks related to gaze following. Our reasoning was that similar underlying mechanisms might be needed to solve these tasks since they require participants to take into account another person's point of view.

The study design and procedure obtained ethical clearance in the same way as Study 1. The study was pre-registered prior to data collection: (https://osf.io/xsqkt). Data collection took place in Leipzig, Germany, between February and March 2023.

Participants

The sample consisted of 102 children (mean age = 4.54 years, SD = 0.31, range = 3.99 - 5.03, 54 girls). Information on individual socio-economic status was not formally recorded.

Procedure

Children were tested in a quiet room in their kindergarten. An experimenter guided the child through the study. For maximum control of extraneous participant variables, we employed a within-subjects study design. Participants performed the tasks in this order: (1) non-social vector following task, (2) ToM task battery, (3) TANGO. We decided on a fixed order to compare participants' performance straight-forwardly with each other. To increase engagement and decrease fatigue or fuzziness, we switched between tablet tasks and tasks with personal interaction. We presented the non-social vector following task before the TANGO so that participants would not be biased to interpret the stimuli as "agent-like".

Non-social vector following. Modeling the structure of the TANGO, we designed a non-social vector following task. This task was presented on a tablet and used the concept of magnetism. On the upper part of the screen, there was a tube with a circular window, containing a gearwheel. On the floor, there was a magnet. The magnet got switched on (with a cartoon-like sound), whereupon the gearwheel moved towards the magnet. The gearwheel moved so that its center aligned with the magnet center while staying inside the circular window. Participants were then asked to locate the magnet. Access to the magnet's true location was manipulated by a wooden wall: participants either had full, partial, or no visual access to the true magnet location. Compared to the TANGO, the circular window acted functionally similar to the agent's eyeball, while the gearwheel acted similar to the pupil. Participants were expected to estimate a vector from the center of the circular window to the gearwheel and extend this as a line toward the ground to locate the magnet. We deliberately decided against displaying an arrow: we aimed to keep the mechanistic functions of the TANGO and magnet stimuli as similar as possible. In both cases, the starting point of the vector needs to be estimated by the participant. With an arrow, we would have drastically reduced the level of uncertainty, since the arrow already displays all information (arrow tip as the "gaze" direction). Furthermore, we wanted to avoid referential or iconic stimuli.

Children received 19 trials with one full visual access trial, two partial visual access trials, and 16 test trials. The first trial of each type comprised a voice-over description of the presented events. We conducted our analysis with 15 test trials (excluding the voice-over trial). The outcome variable was imprecision, defined as the absolute distance between the magnet's x coordinate and the x coordinate of the participant's click. Magnet coordinates were randomized: The full width of the screen was divided into ten bins; each bin occurred equally often, while the same bin could occur in two consecutive trials; and exact coordinates within each bin were randomly generated.

ToM task battery. We administered four tasks from the Wellman and Liu (2004) ToM scale (see Supplements for further detail). We excluded three tasks: the Diverse Desires task to avoid ceiling effects; and both tasks involving emotions (Belief Emotion and Real-Apparent Emotion), as we aimed at assessing the "cold, cognitive" (vs. "emotional") aspects of social cognition. We added two perspective-taking level-2 tasks (Flavell, Flavell, et al. (1981); Flavell, Everett, et al. (1981); where children were asked whether a turtle appeared to be on its back or feet / a worm lay on a red or blue blanket from the experimenter's point of view) with the aim of increasing the variability we can capture between individuals, and since we hypothesized that perspective-taking would rely on similar mechanisms than gaze following, both relying on another's person egocentric frame of reference.

Gaze following. As in Study 1 and 2, we presented children with the TANGO (Prein et al., 2023). To accentuate the social aspect of the TANGO, we exchanged the animal agents (used in the previous two studies) with human faces, which were modeled after the local population in appearance (already created for another project on cross-cultural similarities in gaze following (https://osf.io/tdsvc)). This further highlighted the contrast (i.e., social vs. non-social context) to the non-social vector following task.[7]

Analysis

By design, the TANGO and the non-social vector following task involved vector following. Based on our computational model, we expected children's performance in both tasks to correlate with each other. For each task, we calculated the mean level of imprecision for each subject and correlated them using *Pearson's* correlation coefficient.

For the ToM battery, we aggregated the score of all solved tasks. Please note that the ToM score acted as an umbrella term and included the two perspective-taking tasks. Regarding the relationship between the two vector following tasks and the ToM measures, we could imagine two possible scenarios: (A) If gaze following recruited a social-cognitive ability beyond geometric vector following, we expected that ToM measures would correlate more strongly with the gaze following task than with the non-social vector following task. (B) If gaze following relied purely on task-specific geometric processes, then the correlation between gaze following and ToM measures would be comparable to the correlation between non-social vector following and the ToM measures. For the association between the aggregate ToM scores and the gaze following / non-social vector following tasks, we used *Spearman's* rank correlation coefficients.

---

[7] In an exploratory analysis, we compared children's imprecision levels in the TANGO task with animal vs. human agents. Based on a GLMM analysis, we conclude that there was no evidence of a stable effect of stimulus choice. See Supplements for further detail.

We compared the correlation between gaze following and ToM measures and the correlation between non-social vector following and ToM measures by using the Williams' test from the function `cocor.dep.groups.overlap` (designed for two dependent overlapping correlations) from the package `cocor` (Diedenhofen & Musch, 2015).

To estimate which components best explain the gaze following score, we conducted a model comparison with GLMMs predicting the mean imprecision in gaze following by age, imprecision in non-social vector following, the ToM aggregate score, or the aggregate of the two perspective-taking tasks (subset of ToM battery; example of model notation in `R: tango_mean ~ age_centered + magnet_scaled + perspective_scaled`). The outcome variable was modeled by a lognormal distribution. We wanted to assess whether the ToM aggregate score or the singled-out perspective-taking score added additional explanatory value when predicting the gaze following score. We hypothesized that perspective-taking seemed most closely theoretically related to gaze following as in both cases the participant was asked to judge another person's point of view.

Results

Children performed similarly well in the social and non-social vector following tasks (gaze following mean = 1.92, sd = 0.77, range = [0.57 — 4.49]; vector following mean = 1.90, sd = 1.01, range = [0.57 — 5.83]). The mean aggregate ToM score was 3.46 (sd = 1.53, range = [0 — 6]), while the mean aggregate perspective-taking score was 1.38 (sd = 0.81, range = [0 — 2]).

Gaze following substantially correlated with the non-social vector following task, r = 0.38, 95%CI [0.20, 0.53] (see Figure 4B). While the tasks highly overlap in task demands, their measures shared some, but not all of their variance.

ToM abilities did not correlate with gaze following (ρ = -0.12, 95%CI [-0.31, 0.07]) or non-social vector following (ρ = -0.12, 95%CI [-0.30, 0.08]), and the correlations did not differ from each other, Williams' test $t(99) = 0$, $p = 1$.

Interestingly, gaze following and perspective-taking correlated with each other, ρ = -0.29, 95%CI [-0.46, -0.10]. Please note that the TANGO quantifies imprecision in gaze following. Therefore, a negative correlation suggests that imprecision in gaze following corresponds to less perspective-taking. Non-social vector following and perspective-taking did not correlate, ρ = -0.09, 95%CI [-0.28, 0.10]. However, according to the Williams' test, the two correlations did not differ significantly from each other, $t(99) = -1.86$, $p = 0.07$.

Our model comparison revealed that gaze following was best predicted by a model including non-social vector following (β = 0.14, 95% CrI [0.06; 0.21]) and perspective-taking (β = -0.10; 95% CrI [-0.17, -0.03]), even when controlling for age (β = -0.14, 95% CrI [-0.38, 0.10]) (see Figure 4C and Supplements for the model comparison).

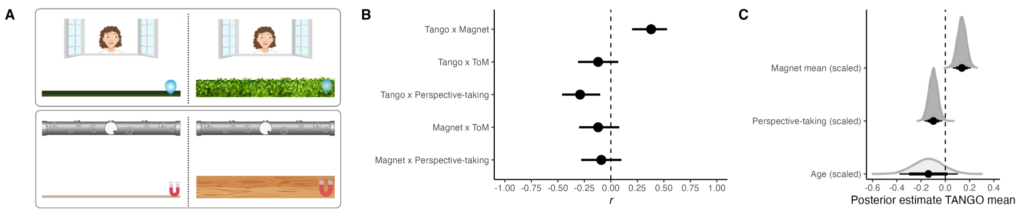

Figure 4: **Components of gaze following.** (A) Study procedures. Top: TANGO (i.e., gaze following; social vector following). Bottom: Magnet (i.e., non-social vector following). The left-hand side shows screenshots of the familiarization phase; right-hand side shows screenshots of the test phase. For illustrative purposes, translucent targets are shown in the test phase. Participants cannot see them. (B) Correlations between gaze following, non-social vector

following, ToM, and perspective-taking. Dots show the correlation coefficients, error bars represent 95% CIs. Please note that the ToM score acts as an umbrella term and also includes two perspective-taking tasks. (C) Influence of age, perspective-taking and non-social vector following on gaze following. The graph shows the posterior distributions for the respective predictor. Black dots represent means, thicker black lines 80% CrI, and thinner black lines 95% CrI.

Discussion

Our gaze model assumed vector calculations on a process level. By isolating non-social vector following experimentally, we could show that gaze following does indeed, to a certain degree, rely on this component. However, non-social vector following alone did not suffice to explain variation in gaze following. Additionally, perspective-taking proved to be a relevant social-cognitive ability when predicting the performance in the TANGO. The overall ToM aggregate score did not add explanatory power.

The TANGO and the two perspective-taking tasks can be seen as instances of visuospatial perspective-taking. Often, researchers distinguish between Level 1 and Level 2 perspective-taking tasks. Level 1 perspective-taking is concerned with the visibility of objects from a particular viewpoint (Surtees et al., 2013), and can usually be solved independently from another's agent frame of reference and without mental rotation [not matching the "mentalizing criterion"; Quesque and Rossetti (2020)]. In contrast, Level 2 perspective-taking tasks ask participants to judge how (visually or conceptually) the world looks for another person. Solving these tasks presumably relies on a simulation of mentally transforming one's own body schema into the space of another agent (Erle & Topolinski, 2017). Participants must understand that different viewpoints lead to different perspectives (Flavell, Everett, et al., 1981): two agents can see the same object "in different, incompatible, fine-grained ways" (Rakoczy, 2022). In short, Level 1 perspective-taking addresses *what* an agent can see, while Level 2 perspective-taking addresses *how or where* they see it (Moll & Tomasello, 2006). The TANGO falls into the first category - however, in contrast to existing research, on a continuous scale - while the perspective-taking tasks presented within the ToM scale fall into the second category.

In the TANGO, participants locate the target by assuming that another agent perceives and looks at the it. The here-applied Level 2 perspective-taking tasks add another layer by asking how one's perspective differs from another's and how exactly the other person sees an object. Michelon and Zacks (2006) assessed the different processes that participants used to master Level 1 and Level 2 perspective-taking tasks and concluded that participants rely on line-of-sight tracing in the former and perspective transformation in the latter case. Line-of-sight tracing consists of locating the avatar and the target, and drawing a line between them. Note how our gaze model in Study 2 shares the underlying idea of connecting points in space via a line. Level 2 perspective-taking requires an additional step in which participants "perform a perspective transformation so that one's imagined position matches the position of the avatar" (Michelon & Zacks, 2006). Therefore, the processes to solve Level 1 perspective-taking tasks might be computationally lighter since there is no need to adapt a new reference frame and there is no potential conflict between one's own and the other person's reference frame. Still, the assumed processes overlap, which could explain the correlation between the TANGO and the administered perspective-taking tasks.

Surtees et al. (2013) further differentiated between visual and spatial perspective-taking. While the former helps to judge if and how an agent sees an object, the latter involves judging the relative spatial locations of an agent and the object. Spatial perspective-taking does not necessitate mental states since computing a line of sight does not demand the presence of another agent (Michelon & Zacks, 2006) and can, therefore, be applied to non-agentive objects with a front (Surtees et al., 2013). This could explain how our participants solved the non-social vector following task.

Interestingly, we found weaker correlations between gaze following and the other ToM tasks. Similarly, in a longitudinal study, Brooks and Meltzoff (2015) found no direct association between gaze following at 10.5 months and explicit ToM at 4.5 years. While the ToM tasks and the TANGO shared the social context, the cognitive processes needed to solve each task might vary. As Rakoczy (2022) reflected, perception-goal psychology (which includes gaze following) comprises understanding that others see different objects or pursue different goals. However, this ability does not necessarily entail understanding more complex meta-representational aspects; for example, understanding that mental states can be false or involve aspectual information.

In previous work, we established that the TANGO is suited to capture meaningful variability across individuals (Prein et al., 2023), which is a crucial task feature when we are interested in revealing the relationship between different cognitive abilities. Importantly, the tasks we used to measure ToM abilities were not designed to capture individual differences: the aggregate score of few dichotomous items is of limited use when it comes to quantifying genuine differences between individuals. However, since these tasks are the gold standard in the social-cognitive literature (Białecka-Pikul et al., 2021; Byom & Mutlu, 2013; Poulin-Dubois et al., 2023; Rakoczy, 2022; Wellman, 2018), and measures with satisfying psychometric properties are, to the best of our knowledge, still scarce (e.g., Beaudoin et al., 2020; Mayes et al., 1996), we nonetheless relied on them in this study. Thus, lower correlations between ToM abilities and gaze following may reflect poor measurement characteristics on the side of ToM tasks rather than a genuine absence of association. We would like to point out that we already stated this concern in our pre-registration (https://osf.io/xsqkt).

If the reliability of the tasks at hand is known, one can estimate the “true” correlation between the latent constructs by applying an attenuation formula or structural equation models (Metsämuuronen, 2022; Trafimow, 2016). Adjusting for the measurement error would increase the so-called true correlation. While we can estimate the split-half reliability for the TANGO (Prein et al., 2023) and the non-social vector following task, we do not have reliability estimates of the ToM measures and cannot apply said approaches. This, in turn, underlines the importance of reporting the psychometric properties of a task. The development of new measures to capture individual differences in social-cognitive abilities seems essential to move this research further.

General discussion

In three studies, we shed light on the cognitive process underlying gaze following and its developmental trajectory across the lifespan. Study 1 focused on how gaze following changes with age. We found a steep performance improvement in the preschool years in which children became more precise in locating the attentional focus of an agent. During teenage years and early adulthood, participants reached their peak performance. Precision levels then stayed comparably stable, with a minor decay toward older adulthood. Beyond these aggregated developmental patterns, we found that individual differences exist throughout the lifespan. In Study 2, we proposed a computational cognitive model that described gaze following at a process level. We modeled gaze following as a process in which participants use the pupil location within the eyes to estimate pupil angles. To locate the attentional focus of an agent (and find a target), they extend the resulting gaze vector towards the ground. Individuals vary in their levels of uncertainty around these pupil angles. Our gaze model outperformed two alternative models, which assumed participants solved the task via a center bias or random guessing. Knowing the TANGO to be a reliable individual differences measure, we investigated potential components of gaze following in Study 3. A fundamental assumption of our computational model was that gaze following relies on a vector following process. We experimentally isolated this component by designing a non-social vector following task. Furthermore, we assessed the relationship between gaze following and traditional ToM tasks. We found that gaze following does, indeed, share a substantial part of its variance with the non-social counterpart of vector following. Additionally, perspective-taking correlated with gaze following, whereas the other ToM measures (focusing on diverse desires, knowledge access, and false beliefs) did not.

The developmental trajectory seen in Study 1 shows how abilities that emerge in infancy can continue to develop throughout childhood. While previous research established that four-month-olds can follow gaze toward one out of two objects, this is not the end point of development. By studying gaze following on a continuum, we assessed not just *whether*, but *how precisely* children locate others' attentional focus. In previous work, we have shown that these individual differences are meaningful (e.g., connected to theoretically related constructs, and showing high split half and retest reliability; (Prein et al., 2023)). Capturing individual variation is crucial when we study development and the improvement in social-cognitive abilities.

Preschool children increased their precision level to locate an agent's attentional focus, which then stayed comparably stable across adulthood. Older adults decreased slightly in their precision levels. This developmental trajectory of a first emergence with a rapid improvement, followed by a plateau and slight decline toward older age, might be representative of many cognitive processes.

In Study 2, we proposed a theoretical framework to interpret the development and individual differences in gaze following. Our computational cognitive model assumes that participants estimate a pupil angle (i.e., the angle between a line extending vertically downwards from the pupil center and a line connecting the pupil and eye center; line-of-sight; see Figure 2A). The model parameter estimates a participant's latent ability to follow gaze and can explain why individuals differ in their precision to locate an agent's attentional focus. The model proposes that development in gaze following equals a reduction in noise when estimating the agent's pupil angles. We found strong evidence for the proposed gaze model when comparing it against two alternative models and correlating its predictions with the observed data (see Figure 3D), both for children and adults. Notably, the model recovers signature patterns in the data (see Figure 3C).

In Study 3, we tested the relation between gaze following and non-social vector following. As implied by our gaze model, we found that gaze following relates to the ability to estimate vectors in space. Already Butterworth and Jarrett (1991) have argued that, as the scene of actions, space is the commonality between different minds. The spatial vector following ability might be helpful in several social-cognitive tasks, for example, action prediction (Friesen & Rao, 2011), and intention understanding. Predicting which object another agent likely wants to grasp or calculating their movement pathway could rely on similar vector following abilities.

Gaze as displayed in the TANGO versus as in real-life social interactions differs with regard to which information is available. In the TANGO, the agent's eyes are big and round, with uniform colors, white sclera, fully visible iris and pupils, and continuous eye movement. Eye gaze in natural social interactions often provides less information and is more ambiguous. Even though the TANGO and our gaze model are designed within a 2D world, we believe the mechanisms can be extrapolated into the 3D world. The processes of understanding gaze in daily life likely rely on the same principles as proposed in this paper. In Study 3, we presented the first evidence that this might be the case. We administered Level 2 perspective-taking tasks in which participants needed to adapt another person's frame of reference in a real-world social interaction. The correlation between this task and the TANGO speaks toward a unified mechanism behind these two visual perspective-taking tasks, regardless of the testing setup or stimulus features. Clearly, our real-life environment is visually more cluttered and diverse than the one presented in our tablet task. Here, however, additional informational sources are available to infer where others are looking; for example, body or head orientation or common ground (Bohn & Köymen, 2018; Moll & Kadipasaoglu, 2013; Osborne-Crowley, 2020). From a modeling perspective, a shared interaction history or diverse desires might be represented as non-uniform prior distributions over locations in the visual scenery. This way, our gaze model could be expanded to include more complex processes like mental state reasoning.

Theories of gaze following differ in whether they illustrate why children pay attention to gaze in the first place versus how they identify the exact location of gaze. While we described the process behind children's increasing precision in gaze following, we still need to further explore the driving forces behind this development. Existing theories broadly vary in how much importance they place on (A) experience and social environment, and (B) social awareness vs. domain-general learning mechanisms (see categorization by Astor and Gredebäck (2022)). For example, it has been hypothesized that infants might be equipped with an innate gaze module and special neural mechanisms to detect eyes (Baron-Cohen, 1995; Batki et al., 2000); that children identify contingencies in social interactions and, in these, get reinforced to follow gaze (Corkum & Moore, 1998; Silverstein et al., 2021; Triesch et al., 2006); or that children are intrinsically socially motivated and simulate their own experiences to understand others (Astor et al., 2020; Friesen & Rao, 2011; Ishikawa et al., 2020; Meltzoff, 2007; Tomasello, 1999). As seen in Prein et al. (2023), precision in gaze following is linked to receptive vocabulary and opportunities for social interaction (e.g., number of siblings and age when entering childcare). Which exact kind of interactions are most helpful to improve precision in gaze following remains unknown.

Limitations

In this paper, we have focused on studying variation across ages and individuals. However, our findings rely on participants from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) backgrounds (Henrich et al., 2010). Cultural variation has been found in many foundational aspects of cognition and socialization: for example, in parent-child interaction and communication (Nielsen et al., 2017). First evidence suggests that cultural variation in face-to-face interactions does not influence infants' gaze shifts (Hernik & Broesch, 2019). While we cannot generalize our here-reported developmental trajectory to different socio-environmental settings, we predict that our presented process-model of gaze following holds true across communities. Analyzing cross-cultural variation and checking for predicted "signature patterns" in the data will inform our modeling work and theory building in further detail (Amir & McAuliffe, 2020).

In Study 1, we recruited older participants online which might have selected a particular subgroup of this age range. Seventy-year-olds who have working Wi-Fi connections, know how to use a computer, and are registered on *Prolific* might not be representative of their age group. We can imagine that results from a more diverse, in-person data collection might show different developmental trajectories toward old age.

Our computational model of gaze following estimates one person-specific parameter for how accurately participants locate another person's attentional focus. The model assumes no motor imprecision in this estimation. However, younger children could have located the agent's focus at one particular point but clicked somewhere slightly off for motor control reasons. This would blur the model's estimation of the inferential component. However, in the first training trial, in which children were simply asked to touch the balloon, we found nearly perfect precision levels (cf. Prein et al., 2023). Motor issues and inaccurate aiming, resulting in falsely wide estimations in the model's inferential component, seem unlikely.

In Study 3, we matched the non-social vector following task as closely as possible to the TANGO. However, the starting positions differ: the magnet never appeared in the center of the screen. The starting point of the balloon might be especially important when interpreting the U-shaped pattern in Study 2. Furthermore, in the TANGO, two eyes are presented and information of the two (matching) cues needs to be integrated to infer the target's location. In the non-social vector following task, only one circular window with a gearwheel inside is presented as a directional cue, and there is no need to integrate two different information sources. In addition, we want to mention that the gaze presented in the TANGO might be more prominent compared to real-life social interactions (e.g., perfectly round and visible sclera; see discussion above). Future research should investigate how factors like self-propelled movement, spatial layout, and number of information sources influence the mechanisms of gaze following.

Conclusion

In three studies, we have illuminated the lifelong development of precisely estimating another's gaze direction, and the psycho-physical process behind this. We have shown that gaze following continues to develop beyond infancy, and that individuals differ in their precision levels to localize the gaze direction of an agent. Our proposed process-level theory of gaze following modeled individual differences in precision as varying levels of uncertainty in the estimated gaze vectors. Consequently, we found that imprecision in gaze following relates to non-social vector following, as proposed by the model. Additionally, gaze following was linked to visual perspective-taking but no other aspects of ToM. The present research shows how precise and reliable measures and process models jointly inform each other and lead to a more comprehensive understanding of the psychological phenomenon in question.

References

Amir, D., & McAuliffe, K. (2020). Cross-cultural, developmental psychology: Integrating approaches and key insights. *Evolution and Human Behavior*, *41*(5), 430–444. https://doi.org/10.1016/j.evolhumbehav.2020.06.006

Anstis, S. (2018). The Role of the Pupil, Corneal Reflex, and Iris in Determining the Perceived Direction of Gaze. *I-Perception*, *9*(4), 2041669518765852. https://doi.org/10.1177/2041669518765852

Anstis, S. M., Mayhew, J. W., & Morley, T. (1969). The Perception of Where a Face or Television 'Portrait' Is Looking. *The American Journal of Psychology*, *82*(4), 474–489. https://doi.org/10.2307/1420441

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278. https://doi.org/10.1037/h0033117

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Astor, K., & Gredebäck, G. (2022). Gaze following in infancy: Five big questions that the field should answer. In *Advances in Child Development and Behavior* (p. S0065240722000192). Elsevier. https://doi.org/10.1016/bs.acdb.2022.04.003

Astor, K., Lindskog, M., Forssman, L., Kenward, B., Fransson, M., Skalkidou, A., Tharner, A., Cassé, J., & 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. (1995). *Mindblindness: An essay on autism and theory of mind*. The MIT Press.

Batki, A., Baron-Cohen, S., Wheelwright, S., Connellan, J., & Ahluwalia, J. (2000). Is there an innate gaze module? Evidence from human neonates. *Infant Behavior and Development*, *23*(2), 223–229. https://doi.org/10.1016/S0163-6383(01)00037-6

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*, *10*, 2905. https://doi.org/10.3389/fpsyg.2019.02905

Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental Science*, *8*(6), 492–499. https://doi.org/10.1111/j.1467-7687.2005.00440.x

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, *78*(2), 161–188. https://doi.org/10.1111/j.1751-5823.2010.00112.x

Białecka-Pikul, M., Białek, A., Kosno, M., Stępień-Nycz, M., Blukacz, M., & Zubek, J. (2021). Early mindreading scale: From joint attention to false-belief understanding. *European Journal of Developmental Psychology*, *0*(0), 1–18. https://doi.org/10.1080/17405629.2021.1911799

Birch, S. A. J., Li, V., Haddock, T., Ghrear, S. E., Brosseau-Liard, P., Baimel, A., & Whyte, M. (2017). Perspectives on Perspective Taking. In *Advances in Child Development and Behavior* (Vol. 52, pp. 185–226). Elsevier. https://doi.org/10.1016/bs.acdb.2016.10.005

Bohn, M., & Frank, M. C. (2019). The Pervasive Role of Pragmatics in Early Language. *Annual Review of Developmental Psychology*, *1*(1), 223–249. https://doi.org/10.1146/annurev-devpsych-121318-085037

Bohn, M., & Köymen, B. (2018). Common Ground and Development. *Child Development Perspectives*, *12*(2), 104–108. https://doi.org/10.1111/cdep.12269

Brooks, R., & Meltzoff, A. N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, *38*(6), 958–966. https://doi.org/10.1037/0012-1649.38.6.958

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology*, *130*, 67–78. https://doi.org/10.1016/j.jecp.2014.09.010

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. 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

Butterworth, G., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, *9*(1), 55–72. https://doi.org/10.1111/j.2044-835X.1991.tb00862.x

Byers-Heinlein, K., Tsui, R. K.-Y., van Renswoude, D., Black, A. K., Barr, R., Brown, A., Colomer, M., Durrant, S., Gampe, A., Gonzalez-Gomez, N., Hay, J. F., Hernik, M., Jartó, M., Kovács, Á. M., Laoun-Rubenstein, A., Lew-Williams, C., Liszkowski, U., Liu, L., Noble, C., … Singh, L. (2021). The development of gaze following in monolingual and bilingual infants: A multi-laboratory study. *Infancy*, *26*(1), 4–38. https://doi.org/10.1111/infa.12360

Byom, L., & Mutlu, B. (2013). Theory of mind: Mechanisms, methods, and new directions. *Frontiers in Human Neuroscience*, *7*.

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4), i–vi, 1–143.

Corkum, V., & Moore, C. (1995). Development of joint visual attention in infants. In *Joint attention: Its origins and role in development* (pp. 61–83). Lawrence Erlbaum Associates, Inc.

Corkum, V., & Moore, C. (1998). The origins of joint visual attention in infants. *Developmental Psychology*, *34*(1), 28–38. https://doi.org/10.1037/0012-1649.34.1.28

D'Entremont, B. (2000). A perceptual–attentional explanation of gaze following in 3- and 6-month-olds. *Developmental Science*, *3*(3), 302–311. https://doi.org/10.1111/1467-7687.00124

D'Entremont, B., Hains, S. M. J., & Muir, D. W. (1997). A demonstration of gaze following in 3- to 6-month-olds. *Infant Behavior and Development*, *20*(4), 569–572. https://doi.org/10.1016/S0163-6383(97)90048-5

Deák, G. O., Flom, R. A., & Pick, A. D. (2000). Effects of gesture and target on 12- and 18-month-olds' joint visual attention to objects in front of or behind them. *Developmental Psychology*, *36*(4), 511–523.

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

Diedenhofen, B., & Musch, J. (2015). Cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, *24*(6), 581–604. https://doi.org/10.1016/S0149-7634(00)00025-7

Erle, T. M., & Topolinski, S. (2017). The grounded nature of psychological perspective-taking. *Journal of Personality and Social Psychology*, *112*(5), 683–695. https://doi.org/10.1037/pspa0000081

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology*, *17*, 99–103. https://doi.org/10.1037/0012-1649.17.1.99

Flavell, J. H., Flavell, E. R., Green, F. L., & Wilcox, S. A. (1981). The Development of Three Spatial Perspective-Taking Rules. *Child Development*, *52*(1), 356–358. https://doi.org/10.2307/1129250

Friesen, A. L., & Rao, R. P. N. (2011). Gaze Following as Goal Inference: A Bayesian Model. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33).

Frith, C. D., & Frith, U. (2012). Mechanisms of Social Cognition. *Annual Review of Psychology*, *63*(1), 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Gamer, M., & Hecht, H. (2007). Are you looking at me? Measuring the cone of gaze. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(3), 705–715. https://doi.org/10.1037/0096-1523.33.3.705

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The Structure of Working Memory From 4 to 15 Years of Age. *Developmental Psychology*, *40*, 177–190. https://doi.org/10.1037/0012-1649.40.2.177

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages*.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, *59*(2), 93–104. https://doi.org/10.1037/0003-066X.59.2.93

Gredebäck, G., Fikke, L., & Melinder, A. (2010). The development of joint visual attention: A longitudinal study of gaze following during interactions with mothers and strangers. *Developmental Science*, *13*(6), 839–848. https://doi.org/10.1111/j.1467-7687.2009.00945.x

Guterstam, A., Kean, H. H., Webb, T. W., Kean, F. S., & Graziano, M. S. A. (2019). Implicit model of other people's visual attention as an invisible, force-carrying beam projecting from the eyes. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(1), 328–333. https://doi.org/10.1073/pnas.1816581115

Heeley, D. W., Buchanan-Smith, H. M., Cromwell, J. A., & Wright, J. S. (1997). The oblique effect in orientation acuity. *Vision Research*, *37*(2), 235–242. https://doi.org/10.1016/S0042-6989(96)00097-1

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2-3), 61-83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

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

Hessels, R. S. (2020). How does gaze to faces support face-to-face interaction? A review and perspective. *Psychonomic Bulletin & Review*, *27*(5), 856–881. https://doi.org/10.3758/s13423-020-01715-w

Horstmann, G., & Linke, L. (2021). Examining Gaze Cone Shape and Size. *Perception*, *50*(12), 1056–1065. https://doi.org/10.1177/03010066211059930

Ishikawa, M., Senju, A., & Itakura, S. (2020). Learning Process of Gaze Following: Computational Modeling Based on Reinforcement Learning. *Frontiers in Psychology*, *11*.

Ishikawa, M., Senju, A., Kato, M., & Itakura, S. (2022). Physiological arousal explains infant gaze following in various social contexts. *Royal Society Open Science*, *9*(8), 220592. https://doi.org/10.1098/rsos.220592

Jasso, H., & Triesch, J. (2006). Using eye direction cues for gaze following – A developmental model. *In*.

Lau, B., & Triesch, J. (2004). Learning gaze following in space: A computational model. *In Proceedings of the Third International Conference on Development and Learning*.

Lempers, J. D. (1979). Young Children's Production and Comprehension of Nonverbal Deictic Behaviors. *The Journal of Genetic Psychology*, *135*(1), 93–102. https://doi.org/10.1080/00221325.1979.10533420

Lempers, J. D., Flavell, E. R., & Flavell, J. H. (1977). The development in very young children of tacit knowledge concerning visual perception. *Genetic Psychology Monographs*, *95*(1), 3–53.

Loomis, J. M., Kelly, J. W., Pusch, M., Bailenson, J. N., & Beall, A. C. (2008). Psychophysics of perceiving eye-gaze and head direction with peripheral vision: Implications for the dynamics of eye-gaze behavior. *Perception*, *37*(9), 1443–1457. https://doi.org/10.1068/p5896

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

Meltzoff, A. N. (2007). “Like me”: A foundation for social cognition. *Developmental Science*, *10*(1), 126–134. https://doi.org/10.1111/j.1467-7687.2007.00574.x

Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. N. (2010). “Social” robots are psychological agents for infants: A test of gaze following. *Neural Networks*, *23*(8), 966–972. https://doi.org/10.1016/j.neunet.2010.09.005

Metsämuuronen, J. (2022). Attenuation-Corrected Estimators of Reliability. *Applied Psychological Measurement*, *46*(8), 720–737. https://doi.org/10.1177/01466216221108131

Michel, C., Kayhan, E., Pauen, S., & Hoehl, S. (2021). Effects of Reinforcement Learning on Gaze Following of Gaze and Head Direction in Early Infancy: An Interactive Eye-Tracking Study. *Child Development*, *92*(4), e364–e382. https://doi.org/10.1111/cdev.13497

Michelon, P., & Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception & Psychophysics*, *68*(2), 327–337. https://doi.org/10.3758/BF03193680

Mikellidou, K., Cicchini, G. M., Thompson, P. G., & Burr, D. C. (2015). The oblique effect is both allocentric and egocentric. *Journal of Vision*, *15*(8), 24. https://doi.org/10.1167/15.8.24

Moll, H., & Kadipasaoglu, D. (2013). The primacy of social over visual perspective-taking. *Frontiers in Human Neuroscience*, *7*.

Moll, H., & Meltzoff, A. (2011). Perspective-Taking and its Foundation in Joint Attention. In J. Roessler, H. Lerman, & N. Eilan (Eds.), *Perception, Causation, and Objectivity* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199692040.003.0016

Moll, H., & Tomasello, M. (2004). 12- and 18-month-old infants follow gaze to spaces behind barriers. *Developmental Science*, *7*(1), F1–F9. https://doi.org/10.1111/j.1467-7687.2004.00315.x

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*(3), 603–613. https://doi.org/10.1348/026151005X55370

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

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Osborne-Crowley, K. (2020). Social Cognition in the Real World: Reconnecting the Study of Social Cognition With Social Reality. *Review of General Psychology*, *24*(2), 144–158. https://doi.org/10.1177/1089268020906483

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

Perner, J., Brandl, J. L., Garnham, A., & Peter Lang. (2003). What is a Perspective Problem? Developmental Issues in Belief Ascription and Dual Identity. *Facta Philosophica*, *5*(2), 355–378. https://doi.org/10.5840/factaphil20035220

Pfeiffer, U. J., Vogeley, K., & Schilbach, L. (2013). From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews*, *37*(10, Part 2), 2516–2528. https://doi.org/10.1016/j.neubiorev.2013.07.017

Poulin-Dubois, D., Goldman, E. J., Meltzer, A., & Psaradellis, E. (2023). Discontinuity from implicit to explicit theory of mind from infancy to preschool age. *Cognitive Development*, *65*, 101273. https://doi.org/10.1016/j.cogdev.2022.101273

Povinelli, D. J., Reaux, J. E., Bierschwale, D. T., Allain, A. D., & Simon, B. B. (1997). Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. *Cognitive Development*, *12*(4), 423–461. https://doi.org/10.1016/S0885-2014(97)90017-4

Prein, J. C., Kalinke, S., Haun, D. B. M., & Bohn, M. (2023). TANGO: A reliable, open-source, browser-based task to assess individual differences in gaze understanding in 3 to 5-year-old children and adults. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02159-5

Quesque, F., & Rossetti, Y. (2020). What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice. *Perspectives on Psychological Science*, *15*(2), 384–396. https://doi.org/10.1177/1745691619896607

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

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, *1*(4), 223–235. https://doi.org/10.1038/s44159-022-00037-z

Recasens, A., Khosla, A., Vondrick, C., & Torralba, A. (2015). Where are they looking? *Advances in Neural Information Processing Systems*, *28*.

Remillard, M. L., Mazor, K. M., Cutrona, S. L., Gurwitz, J. H., & Tjia, J. (2014). Systematic Review of the Use of Online Questionnaires of Older Adults. *Journal of the American Geriatrics Society*, *62*(4), 696–705. https://doi.org/10.1111/jgs.12747

Silverstein, P., Feng, J., Westermann, G., Parise, E., & Twomey, K. E. (2021). Infants Learn to Follow Gaze in Stages: Evidence Confirming a Robotic Prediction. *Open Mind*, *5*, 174–188. https://doi.org/10.1162/opmi_a_00049

Stiefelhagen, R., & Zhu, J. (2002). Head orientation and gaze direction in meetings. *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, 858–859. https://doi.org/10.1145/506443.506634

Surtees, A., Apperly, I., & Samson, D. (2013). Similarities and differences in visual and spatial perspective-taking processes. *Cognition*, *129*(2), 426–438. https://doi.org/10.1016/j.cognition.2013.06.008

Symons, L. A., Lee, K., Cedrone, C. C., & Nishimura, M. (2004). What are you looking at? Acuity for triadic eye gaze. *The Journal of General Psychology*, *131*(4), 451–469.

Todorović, D. (2006). Geometrical basis of perception of gaze direction. *Vision Research*, *46*(21), 3549–3562. https://doi.org/10.1016/j.visres.2006.04.011

Tomasello, M. (1999). Social Cognition Before the Revolution. In *Early Social Cognition*. Psychology Press.

Tomasello, M., & Rakoczy, H. (2003). What Makes Human Cognition Unique? From Individual to Shared to Collective Intentionality. *Mind and Language*, *18*(2), 121–147. https://doi.org/10.1111/1468-0017.00217

Trafimow, D. (2016). The attenuation of correlation coefficients: A statistical literacy issue. *Teaching Statistics*, *38*(1), 25–28. https://doi.org/10.1111/test.12087

Triesch, J., Teuscher, C., Deák, G. O., & Carlson, E. (2006). Gaze following: Why (not) learn it? *Developmental Science*, *9*(2), 125–147. https://doi.org/10.1111/j.1467-7687.2006.00470.x

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Walker, L. D., & Gollin, E. S. (1977). Perspective role-taking in young children. *Journal of Experimental Child Psychology*, *24*(2), 343–357. https://doi.org/10.1016/0022-0965(77)90012-1

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, *15*(6), 728–755. https://doi.org/10.1080/17405629.2018.1435413

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

Yaniv, I., & Shatz, M. (1990). Heuristics of reasoning and analogy in children's visual perspective taking. *Child Development*, *61*(5), 1491–1501. https://doi.org/10.2307/1130758

Zhao, K., Wulder, M. A., Hu, T., Bright, R., Wu, Q., Qin, H., Li, Y., Toman, E., Mallick, B., Zhang, X., & Brown, M. (2019). Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A Bayesian ensemble algorithm. *Remote Sensing of Environment*, *232*, 111181. https://doi.org/10.1016/j.rse.2019.04.034

Zohary, E., Harari, D., Ullman, S., Ben-Zion, I., Doron, R., Attias, S., Porat, Y., Sklar, A. Y., & Mckyton, A. (2022). Gaze following requires early visual experience. *Proceedings of the National Academy of Sciences*, *119*(20), e2117184119. https://doi.org/10.1073/pnas.2117184119

7 Study III

A universal of human social cognition: Children from 17 communities process gaze in similar ways

Manuel Bohn (ORCID: 0000-0001-6006-1348)[1,2,*], Julia Prein (ORCID: 0000-0002-3154-6167)[1,2,*], Agnes Ayikoru[3], Florian M. Bednarski (ORCID: 0000-0003-4384-4791)[4], Ardain Dzabatou[5], Michael C. Frank (ORCID: 0000-0002-7551-4378)[6], Annette M. E. Henderson (ORCID: 0000-0003-4384-4791)[4], Joan Isabella[3], Josefine Kalbitz[2], Patricia Kanngiesser (ORCID:0000-0003-1068-3725)[7], Dilara Keşşafoğlu (ORCID: 0000-0002-7356-0733)[8], Bahar Köymen (ORCID: 0000-0001-5126-8240)[9], Maira V. Manrique-Hernandez[2], Shirley Magazi (ORCID: 0009-0006-0479-9800)[10], Lizbeth Mújica-Manrique[2], Julia Ohlendorf[2], Damilola Olaoba[2], Wesley R. Pieters (ORCID:0000-0002-6152-249X)[10], Sarah Pope-Caldwell[2], Katie Slocombe (ORCID: 0000-0002-7310-1887)[11], Robert Z. Sparks (ORCID: 0000-0001-7545-0522)[6], Jahnavi Sunderarajan[2], Wilson Vieira[2], Zhen Zhang (ORCID: 0000-0001-9300-0920)[12], Yufei Zong (ORCID: 0009-0000-5012-0244)[12], Roman Stengelin (ORCID: 0000-0003-2212-4613)[2,10,+], & Daniel B. M. Haun (ORCID: 0000-0002-3262-645X)[2,+]

[1] Institute of Psychology in Education, Leuphana University Lüneburg

[2] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology

[3] Budongo Conservation Field Station

[4] School of Psychology, University of Auckland

[5] Université Marien Ngouabi

[6] Department of Psychology, Stanford University

[7] School of Psychology, University of Plymouth

[8] Department of Psychology, Koç University

[9] Division of Psychology, Communication, and Human Neuroscience, University of Manchester

[10] Department of Psychology and Social Work, University of Namibia

[11] Department of Psychology, University of York

[12] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences

[*] joint first author

[+] joint last author

Author Note

The authors would like to thank Luke Maurits for statistical advice. Manuel Bohn was supported by a Jacobs Foundation Research Fellowship (2022-1484-00). We are gratful to thank all children and caregivers for participating in the study. We thank the Max Planck Society for the Advancement fo Science.

The authors made the following contributions. Manuel Bohn (ORCID: 0000-0001-6006-1348): Conceptualization, Methodology, Formal Analysis, Writing - Original Draft Preparation, Writing - Review & Editing; Julia Prein (ORCID: 0000-0002-3154-6167): Conceptualization, Methodology, Software, Investigation, Writing - Review & Editing; Agnes Ayikoru: Investigation; Florian M. Bednarski (ORCID: 0000-0003-4384-4791): Investigation, Writing - Review & Editing; Ardain Dzabatou: Investigation; Michael C. Frank (ORCID: 0000-0002-7551-4378): Investigation, Writing - Review & Editing; Annette M. E. Henderson (ORCID: 0000-0003-4384-4791): Investigation, Writing - Review & Editing; Joan Isabella: Investigation; Josefine Kalbitz: Investigation, Writing - Review & Editing; Patricia Kanngiesser (ORCID:0000-0003-1068-3725): Investigation, Writing - Review & Editing; Dilara Keşşafoğlu (ORCID: 0000-0002-7356-0733): Investigation, Writing - Review & Editing; Bahar Köymen (ORCID: 0000-0001-5126-8240): Investigation, Writing - Review & Editing; Maira V. Manrique-Hernandez: Investigation; Shirley Magazi (ORCID: 0009-0006-0479-9800): Investigation; Lizbeth Mújica-Manrique: Investigation, Writing - Review & Editing; Julia Ohlendorf: Investigation; Damilola Olaoba: Investigation; Wesley R. Pieters (ORCID:0000-0002-6152-249X): Investigation, Writing - Review & Editing; Sarah Pope-Caldwell: Investigation; Katie Slocombe (ORCID: 0000-0002-7310-1887): Investigation, Writing - Review & Editing; Robert Z. Sparks (ORCID: 0000-0001-7545-0522): Investigation; Jahnavi Sunderarajan: Investigation; Wilson Vieira: Investigation; Zhen Zhang (ORCID: 0000-0001-9300-0920): Investigation, Writing - Review & Editing; Yufei Zong (ORCID: 0009-0000-5012-0244): Investigation; Roman Stengelin

(ORCID: 0000-0003-2212-4613): Conceptualization, Methodology, Investigation, Writing - Review & Editing; Daniel B. M. Haun (ORCID: 0000-0002-3262-645X): Conceptualization, Funding acquisition, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Manuel Bohn (ORCID: 0000-0001-6006-1348), Universitätsallee 1, 21335 Lüneburg, Germany. E-mail: manuel.bohn@leuphana.de

Abstract

Theoretical accounts assume that key features of human social cognition are universal. Here we focus on gaze-following, the bedrock of social interactions and coordinated activities, to test this claim. In this comprehensive cross-cultural study spanning five continents and 17 distinct cultural communities, we examined the development of gaze-following in early childhood. We identified key processing signatures through a computational model that assumes that participants follow an individual's gaze by estimating a vector emanating from the eye-center through the pupil. Using a single reliable touchscreen-based task, we found these signatures in all communities, suggesting that children worldwide processed gaze in highly similar ways. Absolute differences in performance between groups are accounted for by a cross-culturally consistent relationship between children's exposure to touchscreens and their performance in the task. These results provide strong evidence for a universal process underlying a foundational socio-cognitive ability in humans that can be reliably inferred even in the presence of cultural variation in overt behavior.

Keywords: keywords

Word count: X

A universal of human social cognition: Children from 17 communities process gaze in similar ways

Introduction

Human socio-cognitive skills enable unique forms of communication and cooperation that provide a bedrock for cumulative culture and the formation of complex societies[1–7]. The eyes are the proverbial "window to the mind" and eye gaze is essential for many social reasoning processes[8–10]. Others' eye gaze is used to infer their focus of visual attention, which is a critical aspect of coordinated activities, including communication and cooperation[11–16].

The ability to follow gaze emerges early in development[17–20]. The earliest signs of gaze-following have been found in infants as young as four months[21,22]. Initially, infants rely more on head direction than actual gaze direction[23,24]. Throughout the first two years of life, children refine their abilities: they interpret gaze in mentalistic terms, for example, they follow gaze to locations outside their own visual field by moving around barriers[25]. Importantly, individual differences in children's gaze-following abilities predict later life outcomes, most notably communicative abilities[26]. For example, gaze-following at 10 months predicts language scores at 18 months of age[27]. Difficulties with gaze-following have been linked to developmental disorders, including Autism[28–30]. This work highlights the importance of gaze-following as a foundational building block of human social interaction and its central place in theorizing.

A central assumption in the theoretical and empirical work discussed above is that, despite substantial variation in developmental contexts, gaze-following works and develops in the same way across human societies[31]. This assumption – despite being central to many developmental theories – is currently not supported by evidence. On the contrary, cross-cultural studies have revealed substantial diversity in socio-cognitive development[3,32–35]. One of the very few cross-cultural studies also found differences in the likelihood to follow gaze between communities[36].

One potential source for this paradox lies in the reliance on aggregated measures in cross-cultural studies. Absolute differences in mean performance across communities are interpreted as a signal of different underlying cognitive processes. Here, we resolve this paradox by instead focusing on processing signatures that can be investigated independently of absolute community-level differences. This allows us to directly evaluate the empirical foundation of claims about universal features of human social cognition. To this end, we conducted a pre-registered, large-scale, cross-cultural study on the development of gaze-following abilities to study potentially universal processing signatures.

The processing signatures were derived from a simple computational model that assumes that participants follow gaze by estimating a vector emanating from the eye center through the pupil[37]. The key innovation of the model is that it explains how individuals may use the same cognitive process but still differ in their measured abilities. The process always involves estimating a vector but also involves a degree of uncertainty because the eye center is not directly observable. Individuals are assumed to differ in their level of uncertainty with which they estimate the vector which causes differences in their observable behavior. Importantly, the assumed process leaves a key signature in the data that is observable independent of the absolute level of performance. In the present study, we therefor focus on this signature instead of absolute levels of performance when evaluating the claim whether there is evidence for a universal cognitive mechanism underlying gaze-following.

The 1377 participants who took part in the study lived in 17 different communities across 14 countries and five continents (Fig. 1A, Tab. 1). These countries represent ~46% of the world's population. Communities covered a broad spectrum of geographical locations, social and political systems, languages, and subsistence styles (see Supplemental Materials). This diversity allowed us to overcome the common pitfall of cross-cultural studies that compare urban communities from the global north to rural communities from the global south[38].

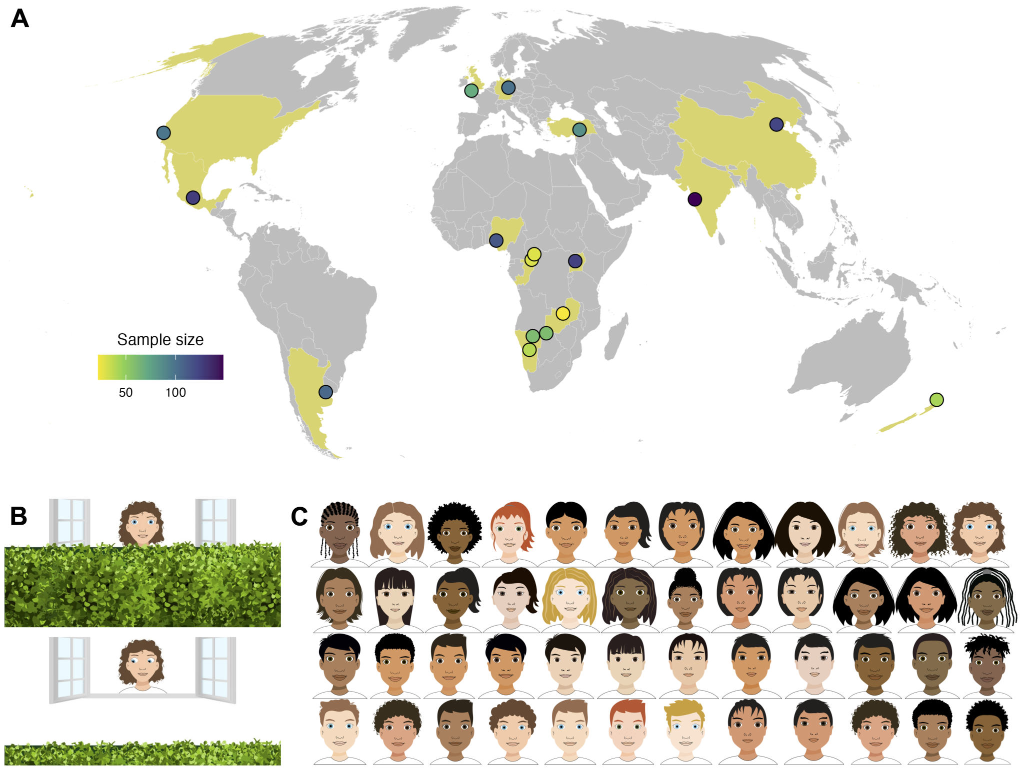

Figure 1. (A) Data collection sites. Points show the approximate geographical location of the data collection sites, coloring shows the sample sizes. (B) Screenshots from the task. Screenshots from the task. The upper scene depicts the start and the lower the choice phase in a test trial. Participants had to use the gaze of the agent to locate the balloon and touch the location on the hedge where they thought the balloon was. Agents, audio recordings and backgrounds were adapted to each cultural setting. (C) Drawings used as agents across cultural settings.

Table 1

Participant demographics.

Continent	Country	Community	N(male)	Age (range)	Language	Touchscreen exposure1
Americas	Argentina	Buenos Aires	105 (53)	4.72 (3.00 - 6.96)	Spanish (Rio-platense)	0.90
	México	Ocuilan	127 (63)	4.96 (2.57 - 6.95)	Spanish (Mexican)	0.77
	USA	Stanford	98 (54)	4.99 (2.52 - 7.90)	English (American)	0.98
Africa	Namibia	Hai||om	60 (38)	5.85 (2.74 - 8.34)	Hai||om	0.05
		Khwe	59 (24)	5.84 (3.38 - 8.63)	Khwedam	0.19
		Windhoek	39 (17)	5.69 (2.66 - 8.66)	English (Nigerian)2	0.95
	Nigeria	Akure	114 (54)	5.07 (2.57 - 7.33)	English (Nigerian)	0.91
	Rep. Congo	BaYaka	29 (13)	7.80 (3.94 - 10.56)	BaYaka	0.00
		Bandongo	30 (11)	7.45 (3.50 - 10.95)	Lingala	0.00
	Uganda	Nyabyeya	125 (62)	5.94 (2.67 - 8.92)	Kiswahili	0.34

Table 1 continued

Continent	Country	Community	N(male)	Age (range)	Language	Touchscreen exposure1
	Zambia	Chimfunshi	22 (5)	5.98 (2.88 - 8.00)	Bemba	0.14
Europe	Germany	Leipzig	100 (48)	4.88 (2.53 - 6.95)	German	0.89
	UK	Plymouth	70 (30)	6.02 (2.38 - 8.94)	English (British)	0.99
Asia	China	Beijing	123 (62)	5.47 (2.69 - 8.48)	Mandarin	0.95
	India	Pune	148 (73)	6.14 (3.06 - 8.83)	English (Indian) / Marathi	0.93
	Türkiye	Malatya	85 (40)	5.02 (2.75 - 7.12)	Turkish	1.00
Oceania	New Zealand	Auckland	43 (19)	5.14 (2.81 - 8.75)	English (New Zealand)	0.95

Note. 1 Proportion of participants who have access to touchscreens according to parental questionnaire. 2 Local collaborators and piloting suggested that Nigerian English is suitable for Windhoek as well.

We used an animated picture book tablet task in which participants had to locate a hidden object based on observing an agent's gaze. Children watched a balloon disappear behind a hedge. An agent followed the trajectory of the balloon with their eyes (Fig. 1B). The key dependent variable was the (im)precision with which children located the agent's focus of attention, that is, the deviation between where the agent looked (where the balloon was) and the child's response. We adapted visuals and audio instructions specifically for each of the 17 communities. Previous work demonstrated excellent individual-level measurement properties for this task in a German sample[39].

Results

Cross-cultural variation in development

As the first step, we investigated developmental improvements, that is, how children become more precise at estimating the target location with age. Across all 17 communities, we found a substantial increase in average levels of precision with age (fixed effect in Bayesian regression model[40]: β = -0.30, 95% Credible Interval (CrI) (-0.40 - -0.21); range of community-level (random) effects: β_{min} = -0.06, 95% CrI (-0.18 - 0.05) to β_{max} = -0.59, 95% CrI (-0.71 - -0.48)).

Nevertheless, there were also marked differences between communities (see Fig. 2A). In a six-fold cross-validation procedure, we trained a regression model on a subset of the data (training data) to later predict the held-out data (testing data)[41]. This procedure was repeated 100 times. We found that a model assuming cross-cultural variation in average performance as well as cross-cultural variation in developmental trajectories outperformed simpler models – assuming no variation in the shape of developmental trajectories or no variation between settings at all – in 98% of cases (see Supplemental Material). There are numerous ways in which communities could be grouped that would fall in line with these absolute differences (e.g., market integration, average levels of education, or average household size). However, we deliberately want to avoid any such overly simplistic explanation based on group-level data. Instead, we think these results can be best understood in methodological terms in the form of exposure to touch-screen devices; a finding we discuss in more detail below. Importantly, average differences in precision between communities were small compared to differences between individuals: communities did not form homogeneous clusters but largely overlapping distributions in that some individuals from communities with a lower average level of precision performed better compared to some individuals from a setting with a very high average level of precision. Similarly, in all communities, some 4-year-olds outperformed children two years older than them (see Fig. 2A). The lack of adequate individual-level measurement instruments in previous large-scale developmental cross-cultural studies made it impossible to contrast these perspectives. The substantial overlap between communities found here speaks against categorical differences in gaze-following and is suggestive of a universal underlying process. However, consistent developmental improvements and overlapping distributions alone cannot inform us about the cognitive processes children use when locating the agent's focus of attention.

Universal processing signatures

Recent computational work modeled gaze-following as social vector estimation[37]. When following gaze, onlookers observer the location of the pupil within the eye and estimate a vector emanating from the center of the eye through the pupil. The focus of attention is the location where the estimated vectors from both eyes hit a surface (Fig. 3). It is assumed that this estimation process has some uncertainty becasue the center of the eye is not directly observable and that individuals vary in their level of uncertainty. As a consequence, even though individuals use the same general process, they might differ in their absolute levels of precision. Crucially, this process model predicts a clear performance signature in our gaze-following task: Trials in which the agent looks further away from the center should result in lower levels of precision compared to trials in which the agent looks closer to the center. This prediction is best understood by considering a similar

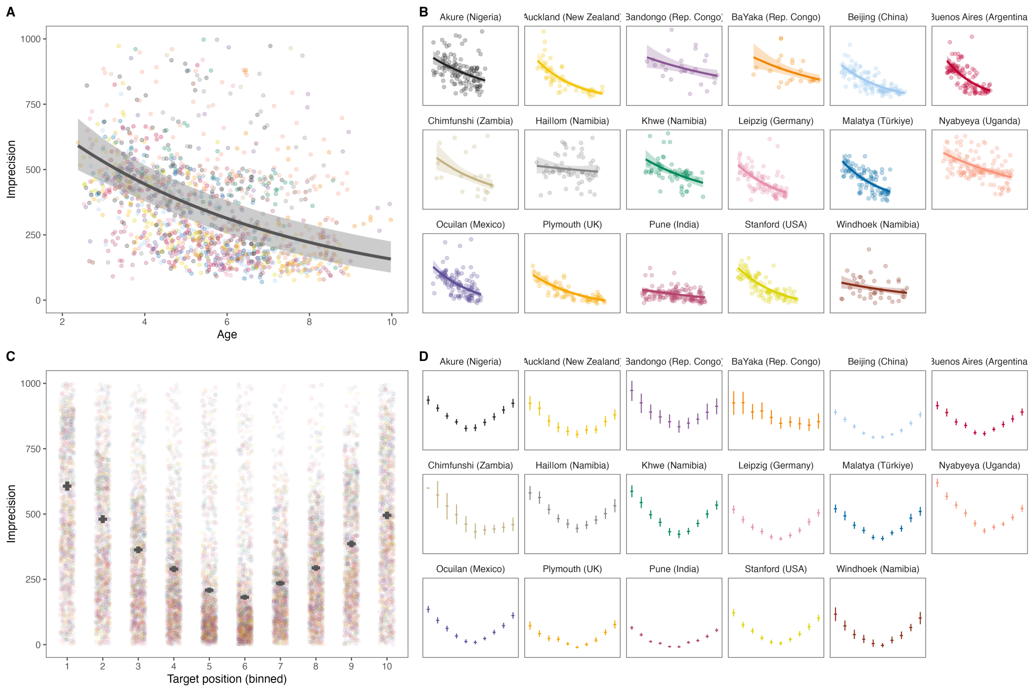

Figure 2. A) Developmental trajectory across and B) by community. The developmental trajectories are predicted based on a model of the data aggregated for each participant. C) Performance by target location on the screen across, and D) by community. Each bin covers 1/10th of the screen. Points show means, and error bars 95% confidence intervals for the data within that bin aggregated across participants. Transparent dots in A) and C) show aggregated data for each individual.

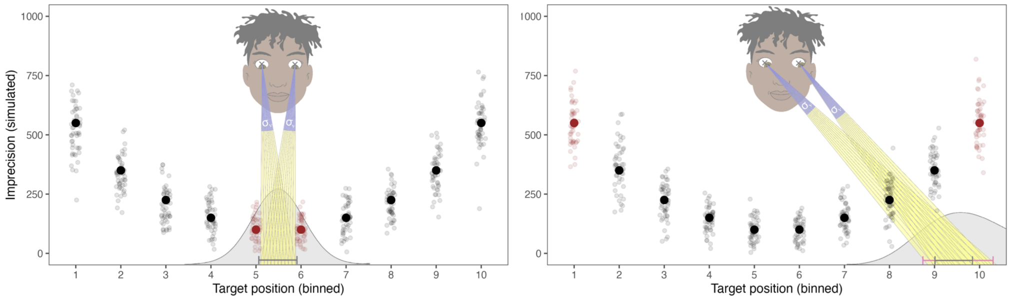

Figure 3. Graphical illustration of the cognitive model. Individuals infer the target of an agent's attention by estimating a vector based on the position of the pupils within the eyes. This process is noisy, illustrated by the different vectors (transparent lines). Individuals differ in their level of precision (indicated by sigma). For a given level of precision, the further the target lands from the centre of the screen, the less precise the model predicts individuals to be. Solid and transparent dots show simulated means and individual data points to illustrate the predicted effect of target position.

phenomenon: pointing a torch light to a flat surface. The width of the light beam represents each individual's level of uncertainty in vector estimation. When the torch is directed straight down, the light beam is concentrated in a relatively small area. When the torch is rotated to the side, the light from one half of the cone must travel further than the light from the other half to reach the surface. As a consequence, the light is spread over a wider area (see Fig. 3).

This processing signature was clearly visible across all 17 communities. Precision decreased when the agent looked at locations further away from the center (fixed effect: β = 0.47, 95% CrI (0.40 - 0.54); range of community-level effects: β_{min} = 0.58, 95% CrI (0.51 - 0.66) to β_{max} = 0.16, 95% CrI (-0.01 - 0.33)). Visualization of the data showed the predicted u-shaped pattern in all communities (see Fig. 2B). These results indicate a universal cognitive process used by children in all communities. There are, however, alternative ways in which the u-shaped pattern might arise: if participants ignored the agent's gaze and instead always selected the middle of the screen (center bias) or randomly selected locations (random guessing), precision would also decrease when the balloon lands further away from the center. To rule out these alternative explanations, we directly compared three models that made different assumptions about how participants' responses were generated: the focal vector-based gaze estimation model described above, a center-bias model where participants always select the center, and a random guessing model where participants select random locations. For every community, we found overwhelming support for the gaze estimation model (min $BF_{10} > 100\,000$ for comparisons with both alternative models). Taken together, children from all 17 communities processed gaze in similar ways.

Predictors of variation

Next, we looked at factors that could explain community- and individual-level variation. In addition to the gaze-following task, caregivers responded to a short questionnaire about children's access to screen-based technology and household composition. On an individual level, we found that children with access to touchscreen devices had higher levels of precision (β = -0.14, SE = 0.04, 95% CrI = -0.21 - -0.07). This effect was consistent across communities in that allowing the effect of access to touchscreens to vary across communities did not improve model fit (see Supplemental Materials). On a community level, we also saw that average performance was lowest in communities in which touchscreen devices were the least frequent (community-level correlation between age-corrected imprecision and proportion of children with access to touchscreens: r = -0.90, 95% CI = -0.96 - -0.74). Thus, familiarity with the device used for data collection likely explains variation between communities. Children with more touchscreen experience were probably better at task handling and thus more likely to precisely touch the location they inferred the agent to look at.

However, there was substantial variation between individuals that could not be explained by differential exposures to touchscreens alone. For example, in Malatya (Türkyie) where 100% of children had access to touchscreens there was still substantial variation between individuals (see Fig.1B). This strongly indicates that other factors likely contributed to individual differences. Social interaction has been highlighted as an important driver of social-cognitive development[e.g., 31,42–45] and thus we hypothesized (and pre-registered) that more opportunities for social interaction – approximated by living in larger households with more children – would be associated with higher levels of precision. When predicting performance by relative opportunities for social interactions within a community – while accounting for absolute differences and the prevalence of touchscreens – we found no strong associations between any of the demographic indicators and performance (see Supplemental Material). Whilst household size was a useful proxy for regular social interaction opportunities, the measure does not directly measure the factors that previous work has suggested to be related to the development of gaze-following in younger children, such as attachment quality or the use of gaze in early communicative interactions[46–48].

Discussion

Following and understanding gaze is a foundational building block of human social cognition[11–16]. A substantial body of work has explored the developmental onset of gaze-following in a few selected cultural communities[17–19,49]. The data reported here provides strong evidence that children from a large and diverse set of communities process others' gaze in similar ways. We found key performance signatures predicted by a model treating gaze-following as a form of social vector estimation across all 17 communities. With the focus on individual-level processing signatures, the study goes beyond previous studies on gaze-following – focused on the onset of gaze-following in infancy[36,50] – as well as comprehensive cross-cultural studies that compared average developmental trajectories[51–54].

The cognitive processes underlying gaze-following might be rooted in humans' evolved cognitive architecture, which is – presumably – later refined during social interaction[46–48]. The phylogenetic roots of these processes might possibly lie much deeper as primates from a wide range of species follow gaze[55–58]. Yet, similarities in overt behavior do not imply the same underlying cognitive processes. The present study defines clear performance signatures that can be explored in other species to test such evolutionary hypotheses.

Our study combined precise individual-level cognitive measurement and individual-level assessment of experience (here: touchscreen exposure) in a large and diverse sample to directly investigate the impact of specific cultural experiences on developmental outcomes. Instead of establishing universality by maximizing the cultural distance between two or three tested communities[59], this large-scale cross-cultural approach treats children's cultural experience at scale, shedding light on the big "middle ground" of children's cultural experience[38].

The study has important limitations. The fact that performance in the task was correlated with exposure to touchscreens might have overshadowed other sources of variation. However, we think it is in an important innovation that we were able to account for this effect. Most developmental cross-cultural studies do not even question the portability of their measurement instruments. Importantly, the key result that the processing signatures were seen in all cultural settings, is immune to this finding. The potential that lies in the otherwise precise individual-level measurement that our task achieves is largely unexploited. The questionnaire items only offer a very coarse picture into children's actual lived experiences. Future work could increase the resolution with which everyday experiences in children from diverse communities are recorded to compare the drivers behind social-cognitive development as we observe it. Recent work in the field of language acquisition has shown how technological innovations allowed for direct recording of social interactions across communities which can be used to close this explanatory gap[60,61].

In sum, our work pioneers an approach that introduces computational modeling and precise individual-level measurement to the cross-cultural study of cognitive development. This approach allowed us to test for universals in the human cognitive architecture rather than just overt behavior. As such, it can serve as a blueprint for future research on a broad spectrum of cognitive abilities and offers a much-needed empirical foundation for theories on the nature of the human mind. Children from diverse cultures deploy similar cognitive processes in interpreting gaze, pointing to a universal foundation of basic social cognition, which is refined during development.

Methods

A more detailed description of the materials and models can be found in the supplemental material. The experimental procedure and analysis plan were pre-registered (https://osf.io/tdsvc). We report on deviations from the pre-registered plan in the supplementary material. The task itself, including all the versions used in the study, can be accessed via the following website: https://ccp-odc.eva.mpg.de/tango-cc/. Data, model and analysis scripts can be found in an online repository (https://github.com/ccp-eva/gafo-cc-analysis).

Participants

A total of 1377 children between 2.38 and 10.95 provided data for the study. Children lived in 17 different communities, located in 14 different countries. Table 1 gives the sample size per community together with some basic demographic information. The recruitment strategy for each community is reported in the respective site descriptions. For some children, the exact birthday was unknown. In such cases, we set the birthday to the 30th of June of the year that would make them fall into the reported age category.

Data from children was only included in the study when they contributed at least four valid test trials. We also excluded the data from children with a diagnosed developmental disorder. In sum, in addition to the sample size reported above, 74 additional children participated in the study but did not contribute data. The main reasons for exclusion were: contribution of less than four valid test trials, technical failures, and missing or implausible demographic information (e.g., when the number of children living in the household was reported to be larger than the household itself or when the number of children reported to live in the household equaled the number of children younger than the child being tested). We did not exclude any participants for performance reasons. A detailed description of each data collection site and the way children were recruited can be found in the supplemental material.

Setup and Procedure

The task was implemented as a browser-based interactive picture book using `HTML` and `JavaScript`. Participants saw animated agents on a touch screen device, listened to pre-recorded audio instructions and responded by touching the screen. In all communities, a research assistant, fluent in the local language(s), guided the child through the task.

Figure 1B shows a screenshot from the task. The task was introduced verbally by the assistant as the balloon game in which the participant would play with other children to find a balloon. On each trial, participants saw an agent located in a window in the center of the screen. A balloon fell down from its starting position just below the agent. The agent's gaze followed the trajectory of the balloon. That is, the pupils and the iris were programmed to align with the center of the balloon. Once the balloon had landed on the ground, the agent was instructed to locate it, that is, to touch the location on the screen where they thought the balloon was. On each trial, we recorded the exact x-coordinate of the participant's touch.

There were two types of familiarization trials. In fam1 trials, the balloon fell down and landed in plain sight. Participants simply had to touch the visible balloon. In fam2 trials, the trajectory of the balloon was visible but it landed behind a small barrier (a hedge - see Figure 1B). Thus, participants needed to touch the hedge where they saw the balloon land. Next came test trials. Here, the barrier moved up and covered the balloon's trajectory. That is, participants only saw the agent's eyes move, but not the balloon. They had to infer the location of the balloon based on the agent's gaze direction. During fam1, fam2 and the first test trial, children heard voice overs commenting what happened on the screen. Critically, the agent was described as wanting to help the child and always looking at the balloon.

Children completed one fam1 trial, two fam2 trials and 16 test trials. We excluded the first test trial from the analysis because of the voice-over. Thus, 15 test trials were used in the analysis below.

Each child saw eight different agents, four male, four female. The agent changed from trial to trial, with alternating genders. A coin toss before the first trial decided whether the first agent was male or female. The order in which agents were shown was randomized with the constraint that all agents had to be shown once until an agent was shown again. The color of the balloon also changed from trial to trial in a random order, also with the constraint that all colors appeared once before any one was repeated.

The location (x-coordinate) where the balloon landed was determined in the following way: The screen was divided in ten equally sized bins. On each trial, one of the bins was randomly selected and the exact x-coordinate was randomly chosen within that bin. Constraints were that the balloon landed in each bin equally often and the same bin appeared no more than twice in a row.

All children were tested with a touchscreen device with a size between 11 and 13 inch equipped with a webcam. The data was either stored locally or sent to a server. In addition to the behavioral data, we stored the webcam recording of the session for verification purposes. Culture-specific adaptations were made by changing the visuals and the audio instructions (see supplementary material for details).

Analysis

We used Bayesian Regression models fit in `R`[62] using the package `brms`[63] for all analyses except the cognitive models. We used the default priors built in to `brms`. The dependent variable in all regression models was imprecision, that is, the absolute distance between the true location of the balloon (x-coordinate of its center) and the location where the participant touched the screen. We used a Log-normal distribution to model the data because the natural lower bound for imprecision is zero and the data was right skewed with a long tail. Numeric predictors that entered the models were scaled to have a mean of zero and a standard deviation of 1.

To analyse cross-cultural variation in performance, we used a cross-validation procedure[see e.g., 41]. In the supplemental material we give a detailed justification of this approach. For each cultural setting, we randomly sampled a data set that was 5/6 the size of the full data set (training data). Then, we fit the model to this training data and used the estimated model parameters to predict the remaining 1/6 of the data (testing data). We then compared the model predictions from the different models by computing the mean difference between the true and predicted imprecision, over all trials in the testing data set. We repeated the cross-validation procedure 100 times and computed the percentage of cases in which one model outperformed the other. We compared three models: a null model assuming no systematic community-level variation, a model assuming variation between communities and a model assuming variation between communities and in developmental trajectories (see supplemental material for model equations).

To evaluate the processing signature that trials in which the balloon lands further away from the center lead to larger imprecision, we fit a model predicting imprecision by age and target centrality (distance of the landing position form the center in pixel) with random intercepts for participant and cultural setting and random slopes for target centrality within participant and cultural setting (`brms` notation: `age + target_centrality + (target_centrality | participant) + (age + target_centrality | culture)`).

Cognitive model

The focal vector-based gaze estimation model has been described in detail in[37]. In brief, it inversely models the process generating touches on the screen based on observed eye movements. Formally, the model is defined as:

$$P(\theta|x_c, \alpha_l, \alpha_r) \propto P(x_c|\alpha_l, \alpha_r, \theta)P(\theta) \tag{1}$$

Here, θ represents an individual's cognitive ability to locate the focus of the agent's attention, x_c represents the touched coordinate, and α_l and α_r correspond to the left and right pupil angles (each defined as the angle between a line connecting the center of the eye to the pupil and a line extended vertically downward from the center of the eye).

The basic assumption in this model is that participants touch on the screen location where they think the agent is looking. The true eye angles (α_l and α_r) are not directly observable and are estimated with noise, yielding $\hat{\alpha_l}$ and $\hat{\alpha_r}$.

Each touch x_c implies a "matched pair" of estimated pupil angles $\hat{\alpha_l}$ and $\hat{\alpha_r}$, with the constraint that the lines extended along those two angles meet at the precise location of where the target is believed to be. As a consequence, we can rewrite the likelihood function of the model as:

$$P(x_c|\alpha_l, \alpha_r, \theta) \propto P(\hat{\alpha_l}, \hat{\alpha_r}|\alpha_l, \alpha_r, \theta)P(x_c) \tag{2}$$

$P(x_c)$ is a prior over potential target locations. Because the target was last visible in the screen and because the agent was located in the center, we assumed that participants have an a priori expectation that the target will land close to the middle. We estimated the strength of this center bias (i.e., the standard deviation of a Normal distribution around the screen center) based on the data: $P(x_c) \sim \mathcal{N}(960, \sigma^p)$.

The primary inferential task for participants is therefore to estimate the pupil angles ($\hat{\alpha_l}$ and $\hat{\alpha_r}$), i.e., to sample from the term $P(\hat{\alpha_l}, \hat{\alpha_r}|\alpha_l, \alpha_r, \theta)$. Here, we assumed that the pair of estimated pupil angles were sampled from a probability distribution which is the product of two Normal distributions of equal variance, σ_v, centered on the true pupil angles:

$$P(\hat{\alpha_l}, \hat{\alpha_r}|\alpha_l, \alpha_r, \theta) \propto \phi(\hat{\alpha}_l; \alpha_l, \sigma_v)\phi(\hat{\alpha}_r; \alpha_r, \sigma_v), \tag{3}$$

Here, σ_v determines the level of accuracy with which participants estimated the pupil angles, and it is thus the component of the model that defines θ. Smaller values of σ_v result in a narrow distribution around the pupil angle, making touches far away from the target less likely. Conversely, larger values for σ_v lead to a wider distribution, making touches far away from the target more likely. To circle back to the analogy introduced above, σ_v corresponds to the width of the light beam. Thus, the goal of the model was to estimate participant-specific values for σ_v: σ_{v_i}. For more details on how σ_{v_i} was estimated, see the supplementary material.

To summarize, the model assumes that participant's touches are generated by a process that relies on noisy estimates of the agent's gaze direction. The precision, with which the gaze direction is estimated, varies between participants and increases with development.

The two alternative models assume that participants ignore the agent's gaze completely, instead they are assumed to follow simple heuristics. According to the center bias model, they always try to touch in the center of the screen: $P(x_c) \sim \mathcal{N}(960, 160)$. (960 is the x-coordinate of the center and 160 is the width of the balloon). According to the random guessing model, they randomly touch coordinates on the screen: $P(x_c) \sim \mathcal{U}(0, 1920)$.

All models were run separately for each cultural setting. The code to run the models can be found in the associated online repository. We also refer to this source for information on the prior distributions for all model parameters.

The cognitive models were implemented in the probabilistic programming language `webppl`[64]. We compared models based on the marginal likelihood of the data for each model, which represents the likelihood of the data while averaging over the prior distribution on parameters. The pair-wise ratio of marginal likelihoods for two models is known as the Bayes Factor. Bayes Factors are a quantitative measure of the predictive quality of a model, taking into account the possible values of the model parameters weighted by their prior probabilities. The incorporation of the prior distribution over parameters in the averaging process implicitly considers model complexity: models with more parameters typically exhibit broader prior distributions over parameter values and broader prior distribution can attenuate the potential gains in predictive accuracy that a model with more parameters might otherwise achieve[65].

References

1. Tomasello, M. The adaptive origins of uniquely human sociality. *Philosophical Transactions of the Royal Society B* **375**, 20190493 (2020).
2. Laland, K. & Seed, A. Understanding human cognitive uniqueness. *Annual Review of Psychology* **72**, 689–716 (2021).
3. Wellman, H. M. *Making minds: How theory of mind develops.* (Oxford University Press, 2014).
4. Henrich, J. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter.* (princeton University press, 2016).
5. Tomasello, M. & Rakoczy, H. What makes human cognition unique? From individual to shared to collective intentionality. *Mind & language* **18**, 121–147 (2003).
6. Legare, C. H. The development of cumulative cultural learning. *Annual Review of Developmental Psychology* **1**, 119–147 (2019).
7. Heyes, C. *Cognitive gadgets.* (Harvard University Press, 2018).
8. Shepherd, S. V. Following gaze: Gaze-following behavior as a window into social cognition. *Frontiers in integrative neuroscience* **4**, 5 (2010).
9. Doherty, M. J. The development of mentalistic gaze understanding. *Infant and Child Development* **15**, 179–186 (2006).
10. Emery, N. J. The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews* **24**, 581–604 (2000).
11. Tomasello, M., Hare, B., Lehmann, H. & Call, J. Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis. *Journal of Human Evolution* **52**, 314–320 (2007).

12. Scaife, M. & Bruner, J. S. The capacity for joint visual attention in the infant. *Nature* **253**, 265–266 (1975).

13. Langton, S. R., Watt, R. J. & Bruce, V. Do the eyes have it? Cues to the direction of social attention. *Trends in cognitive sciences* **4**, 50–59 (2000).

14. Rossano, F. Gaze in conversation. *The handbook of conversation analysis* 308–329 (2012).

15. Richardson, D. C. & Dale, R. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science* **29**, 1045–1060 (2005).

16. Sebanz, N., Bekkering, H. & Knoblich, G. Joint action: Bodies and minds moving together. *Trends in cognitive sciences* **10**, 70–76 (2006).

17. Tang, Y., Gonzalez, M. R. & Deák, G. O. The slow emergence of gaze-and point-following: A longitudinal study of infants from 4 to 12 months. *Developmental Science* e13457 (2023).

18. Gredebäck, G., Fikke, L. & Melinder, A. The development of joint visual attention: A longitudinal study of gaze following during interactions with mothers and strangers. *Developmental science* **13**, 839–848 (2010).

19. Byers-Heinlein, K. *et al.* The development of gaze following in monolingual and bilingual infants: A multi-laboratory study. *Infancy* **26**, 4–38 (2021).

20. Del Bianco, T., Falck-Ytter, T., Thorup, E. & Gredebäck, G. The developmental origins of gaze-following in human infants. *Infancy* **24**, 433–454 (2019).

21. Astor, K., Thiele, M. & Gredebäck, G. Gaze following emergence relies on both perceptual cues and social awareness. *Cognitive Development* **60**, 101121 (2021).

22. D'Entremont, B., Hains, S. M. & Muir, D. W. A demonstration of gaze following in 3-to 6-month-olds. *Infant Behavior and Development* **20**, 569–572 (1997).

23. Michel, C., Kayhan, E., Pauen, S. & Hoehl, S. Effects of reinforcement learning on gaze following of gaze and head direction in early infancy: An interactive eye-tracking study. *Child Development* **92**, e364–e382 (2021).

24. Lempers, J. D., Flavell, E. R. & Flavell, J. H. The development in very young children of tacit knowledge concerning visual perception. *Genetic Psychology Monographs* **95**, 3–53 (1977).

25. Moll, H. & Tomasello, M. 12-and 18-month-old infants follow gaze to spaces behind barriers. *Developmental science* **7**, F1–F9 (2004).

26. Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G. & Moore, C. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development* i–174 (1998).

27. Brooks, R. & Meltzoff, A. N. The development of gaze following and its relation to language. *Developmental science* **8**, 535–543 (2005).

28. Itier, R. J. & Batty, M. Neural bases of eye and gaze processing: The core of social cognition. *Neuroscience & Biobehavioral Reviews* **33**, 843–863 (2009).

29. Thorup, E. *et al.* Altered gaze following during live interaction in infants at risk for autism: An eye tracking study. *Molecular autism* **7**, 1–10 (2016).

30. Thorup, E. *et al.* Reduced alternating gaze during social interaction in infancy is associated with elevated symptoms of autism in toddlerhood. *Journal of Abnormal Child Psychology* **46**, 1547–1561 (2018).

31. Tomasello, M. *Becoming human: A theory of ontogeny.* (Harvard University Press, 2019).

32. Miller, J. G., Wice, M. & Goyal, N. Contributions and challenges of cultural research on the development of social cognition. *Developmental Review* **50**, 65–76 (2018).

33. Mayer, A. & Träuble, B. E. Synchrony in the onset of mental state understanding across cultures? A study among children in samoa. *International Journal of Behavioral Development* **37**, 21–28 (2013).

34. Dixson, H. G., Komugabe-Dixson, A. F., Dixson, B. J. & Low, J. Scaling theory of mind in a small-scale society: A case study from vanuatu. *Child Development* **89**, 2157–2175 (2018).

35. Taumoepeau, M., Sadeghi, S. & Nobilo, A. Cross-cultural differences in children's theory of mind in iran and new zealand: The role of caregiver mental state talk. *Cognitive Development* **51**, 32–45 (2019).

36. Callaghan, T. *et al.* Early social cognition in three cultural contexts. *Monographs of the society for research in child development* i–142 (2011).

37. Prein, J. C., Maurits, L., Werwach, A., Haun, D. B. M. & Bohn, M. Variation in gaze understanding across the life span: A process-level perspective. *PsyArXiv* (2023). doi:10.31234/osf.io/dy73a

38. Barrett, H. C. Towards a cognitive science of the human: Cross-cultural approaches and their urgency. *Trends in Cognitive Sciences* **24**, 620–638 (2020).

39. Prein, J. C., Kalinke, S., Haun, D. B. & Bohn, M. TANGO: A reliable, open-source, browser-based task to assess individual differences in gaze understanding in 3 to 5-year-old children and adults. *Behavior Research Methods* 1–17 (2023).

40. Bürkner, P.-C. Bayesian item response modeling in r with brms and stan. *arXiv preprint arXiv:1905.09501* (2019).

41. Stengelin, R., Ball, R., Maurits, L., Kanngiesser, P. & Haun, D. B. Children over-imitate adults and peers more than puppets. *Developmental Science* **26**, e13303 (2023).

42. Carpendale, J. & Lewis, C. *What makes us human: How minds develop through social interactions.* (Routledge, 2020).

43. Barresi, J. & Moore, C. Intentional relations and social understanding. *Behavioral and brain sciences* **19**, 107–122 (1996).

44. Perner, J., Ruffman, T. & Leekam, S. R. Theory of mind is contagious: You catch it from your sibs. *Child Development* **65**, 1228–1238 (1994).

45. Rakoczy, H. Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology* **1**, 223–235 (2022).

46. Astor, K. *et al.* Social and emotional contexts predict the development of gaze following in early infancy. *Royal Society open science* **7**, 201178 (2020).

47. Movellan, J. R. & Watson, J. S. The development of gaze following as a bayesian systems identification problem. in *Proceedings 2nd international conference on development and learning. ICDL 2002* 34–40 (IEEE, 2002).

48. Senju, A. *et al.* Early social experience affects the development of eye gaze processing. *Current Biology* **25**, 3086–3091 (2015).

49. Moore, C. The development of gaze following. *Child Development Perspectives* **2**, 66–70 (2008).

50. Hernik, M. & Broesch, T. Infant gaze following depends on communicative signals: An eye-tracking study of 5-to 7-month-olds in vanuatu. *Developmental science* **22**, e12779 (2019).

51. House, B. R. *et al.* Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour* **4**, 36–44 (2020).

52. Van Leeuwen, E. J. *et al.* The development of human social learning across seven societies. *Nature Communications* **9**, 2076 (2018).

53. Blake, P. R. *et al.* The ontogeny of fairness in seven societies. *Nature* **528**, 258–261 (2015).

54. Kanngiesser, P. *et al.* Children across societies enforce conventional norms but in culturally variable ways. *Proceedings of the National Academy of Sciences* **119**, e2112521118 (2022).

55. Rosati, A. G. & Hare, B. Looking past the model species: Diversity in gaze-following skills across primates. *Current opinion in neurobiology* **19**, 45–51 (2009).

56. Tomasello, M., Call, J. & Hare, B. Five primate species follow the visual gaze of conspecifics. *Animal behaviour* **55**, 1063–1069 (1998).

57. Itakura, S. Gaze-following and joint visual attention in nonhuman animals. *Japanese Psychological Research* **46**, 216–226 (2004).

58. Kano, F. & Call, J. Cross-species variation in gaze following and conspecific preference among great apes, human infants and adults. *Animal Behaviour* **91**, 137–150 (2014).

59. Norenzayan, A. & Heine, S. J. Psychological universals: What are they and how can we know? *Psychological Bulletin* **131**, 763 (2005).

60. Bergelson, E. *et al.* Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences* **120**, e2300671120 (2023).

61. Donnelly, S. & Kidd, E. The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development* **92**, 609–625 (2021).

62. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, 2023).

63. Bürkner, P.-C. Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software* **80**, 1–28 (2017).

64. Goodman, N. D. & Stuhlmüller, A. The design and implementation of probabilistic programming languages. (2014).

65. Lee, M. D. & Wagenmakers, E.-J. *Bayesian cognitive modeling: A practical course.* (Cambridge University Press, 2014).

8 Study IV

Measuring variation in gaze following across communities, ages, and individuals — a showcase of the TANGO–CC

Abstract

Cross-cultural studies are crucial for investigating the universality and robustness of cognitive developmental processes. Yet, suitable methods to measure variability in cognition across languages and communities are lacking. This paper describes the TANGO–CC (Task for Assessing Individual Differences in Gaze Understanding – Cross-Cultural), a gaze following task designed to measure basic social cognition across individuals, ages, and communities. The TANGO–CC was developed and psychometrically assessed in one setting and subsequently adapted for cross-cultural data collection. Minimal language demands and the web-app implementation allow fast and easy contextual adaptations to each community. The TANGO–CC captured individual differences and showed good internal consistency in a data set from 2.5- to 11-year-old children from 17 diverse communities. Within-community variation outweighed between-community variation. We provide an open-source website for researchers to customize and use the task. The TANGO–CC represents a valuable contribution to assessing basic social cognition in diverse communities, establishing a roadmap for researching cross-cultural individual differences.

Keywords: cross-cultural psychology, social cognition, gaze following, individual differences, reliability

Word count: XXX

Measuring variation in gaze following across communities, ages, and individuals — a showcase of the TANGO–CC

Introduction

For decades, researchers have advocated for more diverse samples in psychological research and cautioned against relying solely on participants from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) backgrounds (Henrich et al., 2010; Lillard, 1998). Despite numerous calls for change, the subject pools reported in high-impact journals still lack diversity (Gutchess & Rajaram, 2023; Nielsen et al., 2017). This lack of representation hinders progress in theory building: inferences about the unique characteristics of human behavior and "human nature" cannot be drawn from humans from only one community (Krys et al., 2024).

To do justice to the diversity in experiences, we must study cognition and its development in diverse communities and even across families and individuals (Gutchess & Rajaram, 2023; Selcuk et al., 2023). So why are not more social-cognitive studies focusing on the variation between individuals living in diverse communities? A potential reason for the under-representation of these studies may be the scarcity of suitable tasks (Bourdage et al., 2023).

Studies investigating variation between communities and/or individuals need to ensure that the captured variation is systematic and not just random noise. This requires measurement reliability and validity. Yet, social cognition studies based on US-American and European samples rarely report psychometric information (for a review, see Beaudoin et al., 2020). This picture further deteriorates when we pay attention to the reliability and validity of cross-cultural social cognition tasks (Bourdage et al., 2023; Hajdúk et al., 2020; Waschl & Chen, 2022). It remains challenging to find reliable and valid tasks that can capture individual differences within one community, let alone tasks that do so across different communities.

Adapting tasks to diverse communities and re-assessing their validity and reliability might be especially important in the social-cognitive domain. If, in theory, stimuli used in social cognition tasks should relate to people's everyday experiences, the tasks need to represent different communities. Indeed, task performance can be diminished when stimuli are not adjusted (Peña, 2007). For example, Elfenbein and Ambady (2002) found better emotion recognition for members of the same national, ethnic, or regional group. Selcuk et al. (2023) concluded that children often attribute mental states more accurately and more frequently to individuals from the same community. This underlines the importance of adapting tasks to each specific cultural context.

Broadly speaking, there are two different approaches that researchers can take to collect cross-cultural data. One approach would be to translate the psychological construct into an individually designed study for each community (termed "assembly"; He and Vijver (2012), Waschl and Chen (2022)). While this approach is most flexible and sensitive to cultural differences, it might be most feasible for studying up to a handful of communities as it becomes too demanding and time-consuming. Most importantly, this approach assumes that the measured underlying concept is the same, while absolute task scores are not comparable across the communities. Another approach would be to use the same standardized procedure across diverse communities, potentially providing a simple translation or modification of culturally inappropriate stimuli (termed "adoption" and "adaptation", respectively; He and Vijver (2012), Waschl and Chen (2022)). This approach is less sensitive to each community's unique characteristics but allows for a direct comparison of the data across communities. Examples following this approach include Callaghan et al. (2011), Taumoepeau et al. (2019), Hughes et al. (2018), Hughes et al. (2014), Fujita et al. (2022), Mehta et al. (2011), Stengelin et al. (2020), and Chasiotis et al. (2006). The present paper aims to describe the development and psychometric properties of a standardized task that can be adapted to diverse communities.

The task presented here focuses on gaze following, that is, the ability to identify the attentional focus of another agent. Gaze following develops early in infancy (Del Bianco et al., 2019; Tang et al., 2024) and contributes to social learning, communication, and collaboration (Bohn & Köymen, 2018; Hernik & Broesch, 2019; Shepherd, 2010; Tomasello et al., 2007). While gaze following is one of the most fundamental social-cognitive abilities, studies focusing on cultural variations are rare. The few existing results are mixed on whether gaze following is influenced by cultural factors or not (Callaghan et al., 2011; Hernik & Broesch, 2019).

The task presented here builds upon the TANGO (Task for Assessing iNdividual differences in Gaze understanding - Open) by Prein et al. (2023). The TANGO measures participants' imprecision in locating an agent's attentional focus. It has been shown to reliably capture individual differences in a German child sample and an English-speaking remote adult sample. The task was sensitive to developmental changes and linked to children's receptive vocabulary. Furthermore, an exploratory analysis showed that children performed equally well in a task version with animal faces compared to cartoon human faces (Prein et al., 2024). This suggests that superficial variations in the stimulus design do not influence children's performance in the task.

This paper showcases the TANGO–CC (TANGO – Cross-Cultural), a standardized gaze following task that can be – and has been – adapted to several languages and communities. We describe the task's development and provide a tutorial for the open-source website (https://ccp-odc.eva.mpg.de/TANGO--CC/). We assess its cross-cultural applicability based on data from a large cross-cultural sample of 2.5- to 11-year-olds from 17 different urban/rural communities across the world and discuss the task's psychometrics. The task and all its adaptations have been initially designed for a paper by Bohn et al. (2024), and we re-use the data set in this paper.

Task development

Approach

In a perfect world, developing a cross-cultural task would include international collaboration and diverse samples from the beginning, already during study design, piloting and item selection. As this seems hardly feasible, we present an alternative, pragmatic approach. First, the task was implemented in one context, in our case with a German sample (Prein et al., 2023). In this context, the task's reliability and validity were assessed in detail. Even though this does not guarantee that the task will be valid and reliable in another context, it substantially increases the likelihood. Second, we reassessed the TANGO–CC's measurement quality (i.e., variability and reliability) across diverse communities by analyzing the data from Bohn et al. (2024), who used the task to collect data in 17 communities. Therefore, our procedure maintains a balance between a detailed analysis of the task's psychometric properties and a swift and feasible task adaptation. In the following, we describe the different steps in further detail. We hope that not just the TANGO–CC but also our pragmatic approach to constructing it will be helpful for other researchers.

In the first step, the task's underlying structure was designed. The TANGO–CC measures the precision with which participants locate an agent's attentional focus. The participant's task is to locate a target by following the agent's gaze (see Figure 1). Precision was measured in a continuous way as the distance between the participant's click on the screen and the target position. The task's core functionality is to animate the agent's eyes so that they follow the target's movement. This basic structure was then embedded in the task's superficial appearance (e.g., background scene) and audio instructions. Once this structure was implemented, adaptations of the task were greatly simplified. For example, we can change the background scene, the faces of the agent, and the target without changing how and what the task measures.

This basic version of the TANGO was psychometrically evaluated in a prototypical WEIRD sample (German child sample; English-speaking remote adult sample) and was found to be highly reliable and valid (Prein et al., 2023). While participants got more and more precise in locating the attentional focus of the agent the older they were, individuals differed across all age groups and showed no floor- or ceiling effects. Performance in the TANGO was linked to children's receptive vocabulary and weakly related to factors of children's daily social environment. In another study, Prein et al. (2024) proposed a computational cognitive model that described gaze following as a social form of vector following. Gaze following, as measured by the TANGO, was related to children's non-social vector following and visual perspective-taking abilities. These connections to related constructs indicate the task's validity in the tested WEIRD setting.

To adapt the task for cross-cultural data collection, we generated a set of human cartoon faces that were judged by researchers and research assistants from each target community to be representative of the local population (see Figure 1). Similarly, different backgrounds were created that roughly represented a typical accommodation in each community. Audio instructions were translated into the corresponding local language. By back-translating these instructions, we ensured the original meaning did not change. Sometimes, specific words were linguistically slightly modified, although functionally equivalent (e.g., “bush” instead of “hedge”), to ensure that all participants understood the instructions. In the following, we describe how researchers can use and customize the TANGO–CC in more detail.

Features of the TANGO–CC

Trials. We quickly recap the TANGO’s (Prein et al., 2023) most characteristic features: Participants are asked to locate a balloon with the help of a gaze cue. The task consists of three different trial types (see Figure 1). In every trial, participants see an agent (boy or girl) looking out of a house with a balloon (red, blue, green, or yellow) in front of them. The balloon falls down to the ground, while the eyes of the agent follow the movement of the balloon in a way that their centers always align. Depending on the trial type, participants have different visual access to the balloon’s position. In training 1, participants see the full trajectory of the balloon and directly have to touch the balloon itself. In training 2, participants see most of the balloon’s movement, but a hedge covers the final location. In test trials, a hedge grows at the beginning of the trial and participants see neither the movement nor the final position of the balloon. The first trial of each type contains an audio description of the presented events (see supplements of Bohn et al. (2024)). Notably, the instructions explicitly state that the agent is looking at the balloon.

The outcome variable is the distance between the participant’s touch and the balloon’s center. Trials can be completed quickly and efficiently so that children can easily complete 15 trials within 10 minutes. This drastically reduces drop-out rates. By using essentially self-explanatory animations, language demands are kept to a minimum. No differential feedback is given to keep trials comparable and avoid learning effects.

Randomization. The order of the agents, balloon colors (red, yellow, green, blue), and balloon positions are each randomized independently. For the balloon positions, the entire width of the screen (1920 in "SVG units") is divided into ten bins. Exact coordinates (value between 0 far left and 1920 far right) within each bin are then randomly generated. The number of repetitions for each agent, balloon color, and balloon bin is calculated based on the total number of trials and the number of unique agents, balloon colors, and bins, respectively. All agents, balloon colors, and bins appear equally often and are not repeated in more than two consecutive trials. If the total number of trials is not divisible by the number of unique elements, additional elements are randomly selected to make up for the remainder.

Cross-cultural customization. The TANGO–CC can be accessed via the following link: (https://ccp-odc.eva.mpg.de/TANGO--CC/). In the first step, researchers can select the language for audio instructions, currently available for 13 different languages and even more dialects (see Table 1). All written instructions are presented in English because they are not directed to the participant but to the research assistant who guides the participant through the task. The task can either be started with the default settings or further customized. The default settings use the version applied in Bohn et al. (2024) based on the selected language.

If researchers choose to customize the task (see Figure 1), the number of trials can be chosen for each trial type. As the trial types build up on each other, each trial type is necessary to understand the structure of the task and needs to be completed before the next trial type starts. Therefore, no trial type can be skipped. The minimum number of trials per type is 1; the maximum is 100. One out of four different backgrounds can be selected. Finally, there are 50 diverse human faces (50% female, 50% male) from which researchers can choose. No constraint exists on how many faces are allowed (min 1, max 50). Once all the settings are adjusted, the customized task is compiled.

In the last step, researchers can enter an alphanumeric participant identifier (1 - 8 characters) and enable a webcam recording of the participant, if needed. To save the selected settings, researchers can bookmark the URL so that the customized task can be easily accessed, and only the participant ID and choice of webcam recording need to be entered again. The task can then be started.

The source code of the task is available on GitHub (https://github.com/ccp-eva/TANGO--CC). By directly editing the `HTML` and `JavaScript` code, researchers gain even more flexibility in adjusting the task to their needs.

Table

Languages available for the audio instructions in the TANGO–CC

Languages	Language family	Speaker's country of origin
Bemba	Bantu	Zambia
Chinese	Sino-Tibetan	China
English	Indo-European	USA / UK / India / Nigeria / New Zealand
German	Indo-European	Germany
Hai||om	Khoesan	Namibia
Khewdam	Khoesan	Namibia
Lingala	Bantu	Rep. Congo
Marathi	Indo-European	India
Shona	Bantu	Zimbabwe
Spanish	Indo-European	Argentina / Mexico

Languages	Language family	Speaker's country of origin
Swahili	Bantu	Uganda
Turkish	Turkic	Türkiye
Yaka	Bantu	Rep. Congo

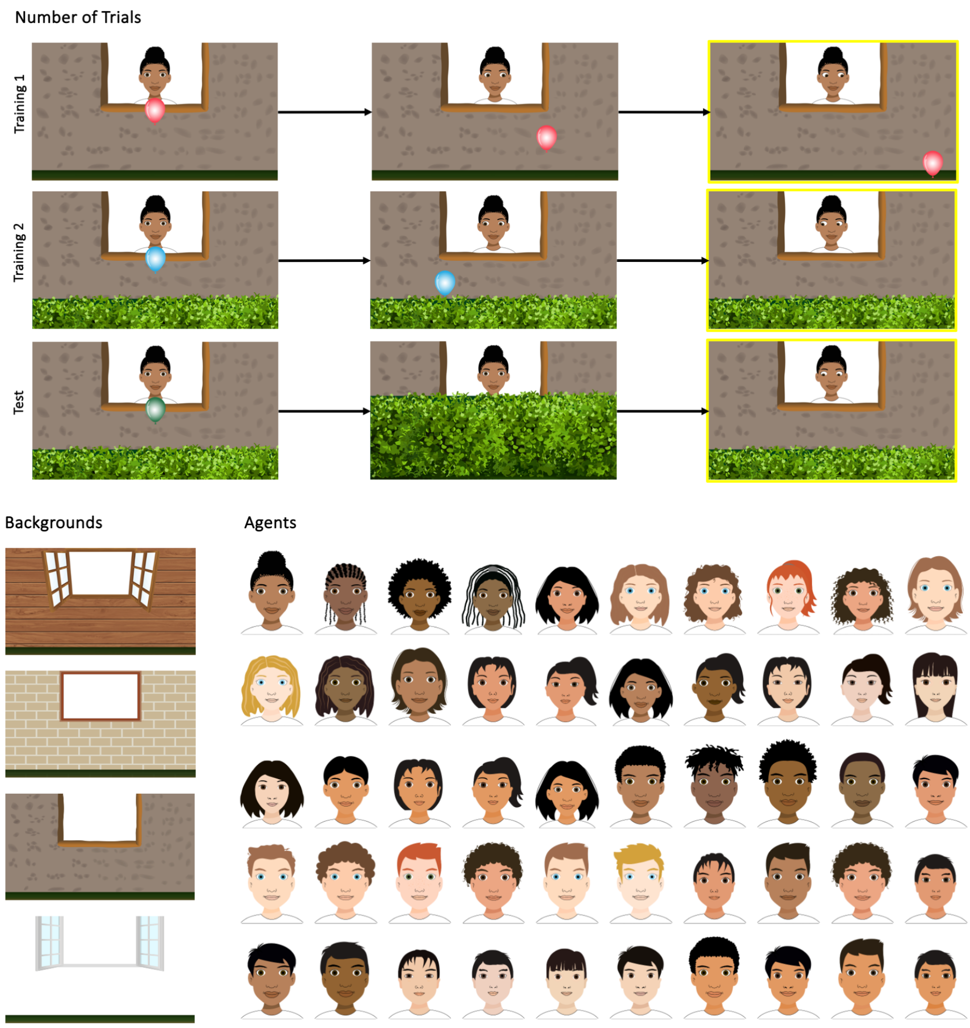

Figure 1: **Customizable components of the TANGO–CC.** Researchers can select the language of the audio instructions, the number of trials per trial type, the background, and the agent's face. Screenshots of the trials show the proceeding events: In training 1, an agent looks at a balloon that falls to the ground, and participants have to respond by touching the balloon. In training 2, the balloon falls behind the hedge while its flight is still visible. Participants respond by touching the hedge where they think the balloon is. In test trials, the balloon's movement and final position are covered by a hedge, and participants respond by touching the hedge. In the task, all movements are smoothly animated (no still pictures). Yellow frames indicate the time point when participants respond (only illustrative, not shown during the task).

Task implementation. The task was implemented in `JavaScript`, `HTML`, and `CSS` and is presented as a web app. It can be accessed on any web browser and does not require prior installation. The online version of the task has been proven convenient for unsupervised data collection (for example, using participant recruitment services like *Prolific*; see Prein et al. (2023)) and sharing the task internationally. Importantly, the web app implementation does not necessarily need a working WIFI connection: An offline, local version of the task can be quickly set up for devices that support Node.js (https://nodejs.org/en). This is an especially useful feature for researchers working in remote areas with limited internet access.

The stimuli are embedded as Scalable Vector Graphics (SVG). The setup allows for an easy adaptation of task elements and ensures that picture quality, aspect ratio, and relative object positioning are constant. The task is programmed so that responses are only registered when the participant touches the relevant part of the screen (i.e., in test trials, the hedge). Furthermore, clicks are only registered after the voice recordings stop playing. An audio reminder is played again if no click is registered within 5 seconds.

The website does not use cookies, nor does it upload any data to servers; that is, the data is only stored locally on the device. The output of the task is a CSV file (and WEBM file if a webcam recording was selected) that contains the participants' responses and can be easily imported into statistical software for further analysis. The file will be stored in the device's downloads folder and is named after the following pattern: "tangoCC-participantID-YYYY-MM-DD_hh_mm_ss".

Psychometric evaluation

Data set

We used the data set from Bohn et al. (2024) for the psychometric evaluation of the TANGO–CC. They collected data using the TANGO–CC in a sample of *N* = 1377 children between 2.5 to 11 years of age. Participants came from 17 communities on five continents, in rural and urban settings, with varying degrees of market integration and technology exposure. Bohn et al. (2024) carried out 19 trials (1 training 1, 2 training 2, and 16 test trials, of which the first of each type had audio instructions). Faces, backgrounds, and languages were chosen by researchers and assistants with experience in the specific community. For further details on the communities and data collection procedures, see the supplements of Bohn et al. (2024).

Individual differences

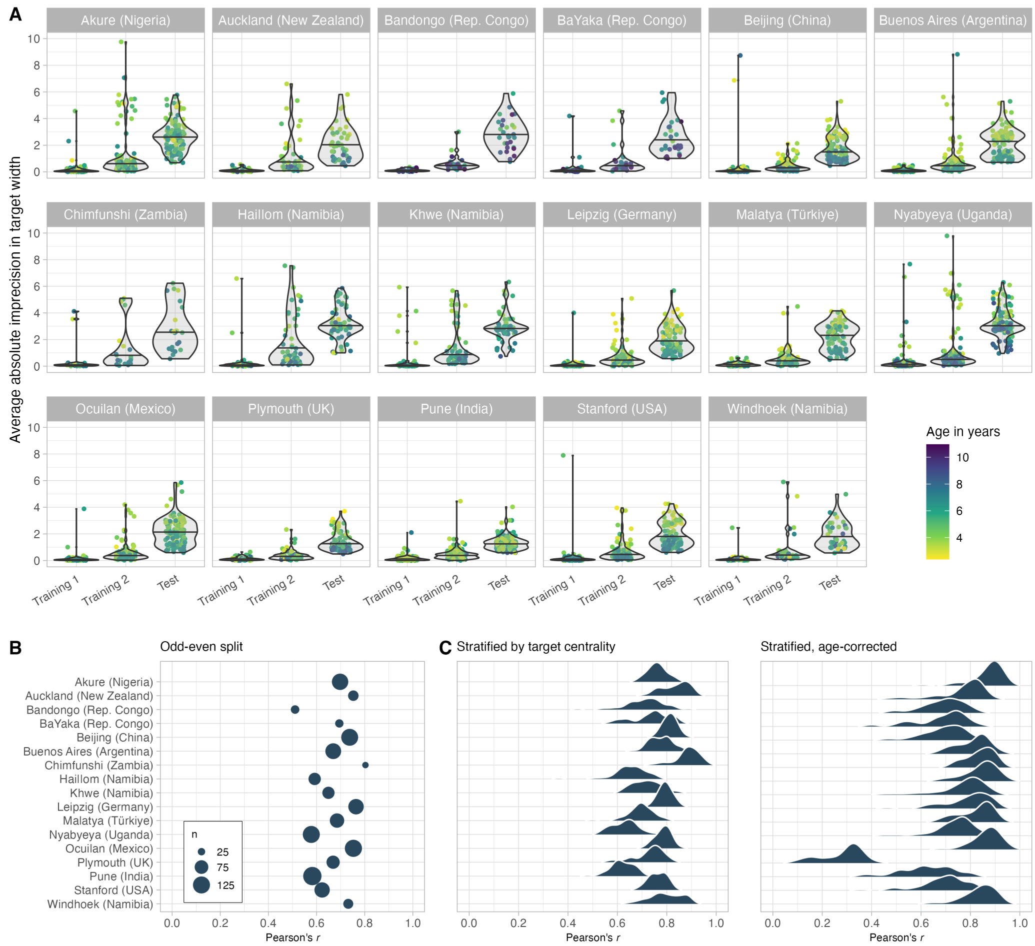

Figure 2: **Measurement of the TANGO–CC by community.** (A) Mean imprecision in locating the agent's attentional focus by community (alphabetically) and trial type. Imprecision is defined as the distance between the participant's touch and the balloon's center in units of balloon width. For a depiction of each trial's procedure, see Figure 1. (B) Internal consistency estimates by community, following three different approaches. In the odd-even split, the size of points reflects the sample size in each community. In the stratified approach with and without age correction, density curves show the posterior distributions of the GLMM.

As a first feasibility check, we inspected the mean and standard deviations by community and compared performance in each trial type (training 1, training 2, test trials). Performance was defined as the absolute click distance between the target center and the click x coordinate, scaled according to balloon widths. Across communities, children performed best in training 1 (mean = 0.19, sd = 0.63), followed by training 2 (mean = 0.79, sd = 1.44) and test trials (mean = 2.21, sd = 2.03; see Figure 2A).

To formally estimate the effect of trial type on performance in the TANGO–CC, we fit a generalized linear mixed model (GLMM) predicting the task performance by trial type (reference category: test trials). All analyses were run in R version 4.3.3 (2024-02-29) (R Core Team, 2024). GLMMs were fit with default priors using the function `brm` from the package `brms` (Bürkner, 2017, 2018). The model included random effects for trial type by community (model notation in `R: imprecision ~ trialtype + (trialtype | community)`), and imprecision was modeled by a `lognormal` distribution. We inspected the posterior distribution (mean and 95% Credible Interval (CrI)) for the trial type estimates.

Our GLMM analysis supported the visual inspection of the data: the estimates for training 1 (β = -3.26; 95% CrI [-3.41; -3.10]) and training 2 (β = -1.47; 95% CrI [-1.58; -1.35]) were negative and reliably different from zero. Please note that the TANGO–CC measures imprecision in gaze following. Therefore, a negative sign shows that children showed less imprecision (i.e., were more precise) in the training trials than in the test trials. This effect was found across all communities (minimum estimate for training 1 = -2.87; minimum estimate for training 2 = -1.27). The almost perfect performance in training trials indicated that children understood the task and were able to locate the balloon. In test trials, children's imprecision was higher, indicating that the task was more challenging. All communities showed great individual variation and overlapped in their imprecision levels (see Figure 2A).

To identify the sources of variation, we computed intraclass correlations (ICC). The variation of children within communities was substantially larger than the variation between the communities. The mean within-community variance was 1.28, ranging from 0.24 (in Pune, India) to 3.46 (in Chimfunshi, Zambia). Between-community variance was 0.34. The ICC, representing the proportion of between-community variance relative to the total variance (sum of within- and between-community variance), was 0.02. This indicates that only 2% of the total variability in the data can be attributed to differences between communities, while the remaining 98% are attributed to differences within communities (Kusano et al., 2024).

Reliability

To assess reliability, we estimated internal consistency in each community in three different ways. First, data of each participant was split into odd and even trials and a Pearson correlation was calculated between the aggregated scores of the two halves. Second, using the function `by_split` from the `splithalfr` package (Pronk et al., 2022), data was stratified by target centrality (capturing trial difficulty), and a Pearson correlation was calculated between the matched halves. Third, a data set was generated with stratified test halves by target centrality. Then, we followed the Generalized Linear Mixed Model (GLMM) approach introduced by Rouder and Haaf (2019). A GLMM was fitted with the mean imprecision as the outcome, age as the predictor, and test half and participant id as random effects (model notation: `imprecision ~ age + (0 + half | subjid)`). The model estimates correlations between participant-specific estimates for each test half. The hierarchical shrinkage of the model enables accurate person-specific estimates. By incorporating age as a fixed effect, the correlation between the two person-specific estimates represents the age-independent estimate for internal consistency. This eliminates the chance that a good internal consistency estimate results from general cognitive development rather than task-specific inter-individual differences. Because the process of generating stratified data sets is partly random, the model was fit 50 times for each community. The posterior estimate of the correlation between the two person-specific estimates was taken as the age-independent estimate for internal consistency.

The results are shown in Figure 2C. Across communities, internal consistency estimates ranged from 0.51 to 0.80 for the odd-even split, 0.62 to 0.89 for the stratified internal consistency, and 0.62 to 0.87 for the age-corrected approach (Plymouth, UK, being an outlier with 0.28). Following Cohen's suggestions (Cohen, 1988, 1992), these correlations constitute large effects (r >.50), and indicated good internal consistency.[1] The results are comparable to the

[1] Note that for scale reliability and Cronbach's α, values of .7 to .8 have been suggested to be acceptable (Field et al., 2012; Kline, 1999). However, Kline (1999) suggested that values below .7 could be realistic for psychological constructs due to their variable nature.

internal consistency estimates found in the original TANGO study (Prein et al., 2023), and also resemble reliability estimates of classical false belief tasks (Hughes et al., 2000).

In an exploratory analysis, we found that communities with larger individual variation showed higher internal consistency estimates (Pearson's r = 0.46, 95%CI [-0.03; 0.77]). Please note that this could be influenced by outliers and that the sample size here (N = 17 communities) is too small to make substantial claims.

Discussion

The TANGO–CC measures imprecision in gaze following across individuals, ages, and communities. The task was developed in two phases. First, the task's underlying functionality was designed in one community. Next, we adapted the superficial features of the task to be used in 17 diverse communities and assessed the task's psychometric properties. Children's imprecision in gaze following highly overlapped between communities: children performed similarly in the communities depending on the trial type, and within-community variation greatly exceeded between-community variation. The task showed satisfactory to high reliability across all communities. Therefore, we believe the TANGO–CC is a promising task to capture individual differences in social-cognitive development in diverse communities. Its design process lays out a much-needed pragmatic approach to conducting cross-cultural individual differences research.

A similar approach to task development was taken by Mehta et al. (2011). The researchers (1) selected social cognition measures that have been established in WEIRD settings, (2) adapted the agent's names, appearance, backgrounds, and languages to the local context, and (3) assessed the task's validity and internal consistency in the new setting. Participants were adults with and without schizophrenia from India. Theory of Mind tasks included Sally-Anne, Smarties, Ice cream van and Missing cookies stories. For this specific context, the authors' approach yielded a successful adaptation of social cognition measures for the tested Indian (Hindi/Kannada) communities.

Bourdage et al. (2023) pointed out a major challenge with adapting social cognition tasks to diverse communities: the number of world cultures is vast, and communities are constantly changing. Therefore, a promising approach might be to provide tasks with a modular system where components can be exchanged according to the local context. In the case of the TANGO–CC, the task can not only be adapted to different languages, cartoon faces, and backgrounds (see Figure 1) but also updated with new stimuli. Unlike studies that present sequential, hand-painted pictures that are difficult to adapt (Mehta et al., 2011), the TANGO–CC uses SVGs that can be easily exchanged.

The biggest strength of the TANGO–CC is its flexibility. The task is presented as a web app that can also run offline to enable remote data collection. Minimal language demands and an engaging, playful design increase the task's usability. Together with a short task duration, this reduces drop-out rates and enables efficient data collection with large sample sizes. The TANGO–CC follows a standardized procedure and uses a continuous, objective outcome measure (leaving no room for rater errors). An online manual with the most frequently asked questions is available at https://ccp-odc.eva.mpg.de/TANGO--CC/manual.html. Additional customization can be achieved by adding new stimuli to the open-source code available on GitHub (https://github.com/ccp-eva/TANGO--CC).

For years, researchers have called for more diverse sampling and culturally valid measures of cognitive development (Matsumoto & Yoo, 2006; e.g., Mehta et al., 2011; Nielsen et al., 2017). As Hajdúk et al. (2020) put it, "using large samples and multisite approaches will align with efforts to improve reproducibility and will clarify both the type and extent of cultural influences on social cognition" (p. 463). The TANGO–CC takes a valuable step in this direction by sharing the task and its source code with other researchers. Bohn et al. (2024) showed that data collection with the TANGO–CC was feasible in 17 diverse communities in rural and urban settings with varying degrees of market integration and technology exposure. While we cannot generalize our findings to all communities worldwide, we found that it induced reliable individual variation in the 17 communities studied by Bohn et al. (2024). Using the TANGO–CC in a new community nevertheless requires sensitivity to the specific context, piloting, and, most importantly, the involvement of researchers or research assistants from the specific community. We hope that the TANGO–CC will facilitate future cross-cultural studies to assess social-cognitive development in a wide range of communities.

A valid question is whether the TANGO–CC measures the same construct across different groups. This so-called measurement invariance is often seen as a requirement for a "fair" cross-cultural comparison and relies on minimizing group differences while individual differences are magnified. As Kusano et al. (2024) put it: "The research challenge is to achieve a balance between ensuring methodological "fairness" at the individual level while also recognizing and capturing genuine sociocultural variability" (p. 34). We argue that the TANGO–CC measures a fundamental social-cognitive ability that is likely similar across communities. Bohn et al. (2024) have shown that children with no prior touchscreen exposure were less precise in the TANGO–CC than children with prior experience. However, individual differences were also found in communities with 100% touch screen exposure, showing that this factor alone could not explain children's performance in the task (Bohn et al., 2024). Notably, even though the touchscreen experience caused absolute differences in task performance, all communities showed the same processing signature. A recent computational cognitive model described gaze following as a process of estimating pupil angles and the corresponding gaze vectors (Prein et al., 2024). The model predicted that all individuals use the same process to locate an attentional focus but differ in their uncertainty around the estimated pupil angles, which results in less precision. Bohn et al. (2024) found clear support for this model in every community they studied, suggesting that children all over the world process gaze in a similar way. In this manuscript, we could also show that internal consistency was high across all communities, meaning that the task captured individual differences in a similar way. Consequently, the TANGO–CC seems to measure systematic individual differences across diverse communities.

Selcuk et al. (2023) pointed out that researchers should study both within- and between-culture variability in the development of social cognition since sometimes within-culture differences exceed between-culture differences. Indeed, we found that within-group variability was greater than between-group variability. While we believe that the TANGO–CC can be used to compare mean differences across communities, we would recommend using it to study individual differences within communities.

Limitations

The TANGO–CC and its psychometric properties need to be considered against some limitations. Reliability for each community was assessed by calculating the internal consistency. Ideally, we would have additionally assessed the task's retest reliability in each community and checked for relationships with theoretically related constructs to assess validity.

Schilbach et al. (2013) pointed out that witnessing social interactions as an observer undoubtedly differs from actively participating in social interactions. Of course, the TANGO–CC does not depict a real-life social interaction, and future research should investigate how task performance relates to the real world. Suggestive evidence comes from a study by Prein et al. (2024), who found that children's performance in the TANGO was linked to children's visual perspective-taking abilities in real-life social interaction. The mode of stimulus presentation surely needs to be kept in mind when administering the TANGO–CC, especially in communities with little technology exposure. Additional touch screen training (e.g., more trials of training 1) might prove helpful in these cases.

Conclusion

The TANGO–CC is a promising task to capture individual differences in social-cognitive development across diverse communities. The task was developed in two phases: (1) implementing the task's underlying functionality and estimating detailed psychometrics in one community, and (2) expanding the stimulus pool to accommodate diverse communities worldwide. The task's flexibility, minimal language demands, and engaging design make it a valuable task for cross-cultural research. The task showed satisfactory to high reliability (internal consistency) in a large dataset including 17 diverse communities. We hope that the TANGO–CC – and its pragmatic construction process – will inspire future cross-cultural studies to assess cognitive development in a wide range of communities.

References

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*, *10*, 2905. https://doi.org/10.3389/fpsyg.2019.02905

Bohn, M., & Köymen, B. (2018). Common Ground and Development. *Child Development Perspectives*, *12*(2), 104–108. https://doi.org/10.1111/cdep.12269

Bohn, M., Prein, J. C., Ayikoru, A., Bednarski, F. M., Dzabatou, A., Frank, M. C., Henderson, A. M. E., Isabella, J., Kalbitz, J., Kanngiesser, P., Keşşafoğlu, D., Koymen, B., Manrique-Hernandez, M., Magazi, S., Mújica-Manrique, L., Ohlendorf, J., Olaoba, D., Pieters, W., Pope-Caldwell, S., … Haun, D. (2024). *A universal of human social cognition: Children from 17 communities process gaze in similar ways*. OSF. https://doi.org/10.31234/osf.io/z3ahv

Bourdage, R., Narme, P., Neeskens, R., Papma, J., & Franzen, S. (2023). An Evaluation of Cross-Cultural Adaptations of Social Cognition Testing: A Systematic Review. *Neuropsychology Review*. https://doi.org/10.1007/s11065-023-09616-0

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. 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

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., & Tomasello, M. (2011). Early social cognition in three cultural contexts. *Monographs of the Society for*

Research in Child Development, *76*(2), vii–viii, 1–142. https://doi.org/10.1111/j.1540-5834.2011.00603.x

Chasiotis, A., Kiessling, F., Hofer, J., & Campos, D. (2006). Theory of mind and inhibitory control in three cultures: Conflict inhibition predicts false belief understanding in Germany, Costa Rica and Cameroon. *International Journal of Behavioral Development*, *30*(3), 249–260. https://doi.org/10.1177/0165025406066759

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

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

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203–235. https://doi.org/10.1037/0033-2909.128.2.203

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.

Fujita, N., Devine, R. T., & Hughes, C. (2022). Theory of mind and executive function in early childhood: A cross-cultural investigation. *Cognitive Development*, *61*, 101150. https://doi.org/10.1016/j.cogdev.2021.101150

Gutchess, A., & Rajaram, S. (2023). Consideration of culture in cognition: How we can enrich methodology and theory. *Psychonomic Bulletin & Review*, *30*(3), 914–931. https://doi.org/10.3758/s13423-022-02227-5

Hajdúk, M., Achim, A. M., Brunet – Gouet, E., Mehta, U. M., & Pinkham, A. E. (2020). How to move forward in social cognition research? Put it into an international perspective. *Schizophrenia Research*, *215*, 463–464. https://doi.org/10.1016/j.schres.2019.10.001

He, J., & Vijver, F. van de. (2012). Bias and Equivalence in Cross-Cultural Research. *Online Readings in Psychology and Culture*, *2*(2). https://doi.org/10.9707/2307-0919.1111

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2-3), 61-83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

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

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., Ensor, R., Koyasu, M., Mizokawa, A., & Lecce, S. (2014). Lost in Translation? Comparing British, Japanese, and Italian Children's Theory-of-Mind

Performance. *Child Development Research*, *2014*, e893492. https://doi.org/10.1155/2014/893492

Hughes, C., Devine, R. T., & Wang, Z. (2018). Does Parental Mind-Mindedness Account for Cross-Cultural Differences in Preschoolers' Theory of Mind? *Child Development*, *89*(4), 1296–1310. https://doi.org/10.1111/cdev.12746

Kline, P. (1999). *The Handbook of Psychological Testing* (2nd ed.). Routledge.

Krys, K., De Almeida, I., Wasiel, A., & Vignoles, V. L. (2024). WEIRD–Confucian comparisons: Ongoing cultural biases in psychology's evidence base and some recommendations for improving global representation. *American Psychologist*. https://doi.org/10.1037/amp0001298

Kusano, K., Napier, J., & Jost, J. (2024). *The Mismeasure of Culture: When Measurement Invariance Requirements Hinder Cross-Cultural Research in Psychology*. OSF. https://doi.org/10.31234/osf.io/9qe2k

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin*, *123*(1), 3–32. https://doi.org/10.1037/0033-2909.123.1.3

Matsumoto, D., & Yoo, S. H. (2006). Toward a New Generation of Cross-Cultural Research. *Perspectives on Psychological Science*, *1*(3), 234–250. https://doi.org/10.1111/j.1745-6916.2006.00014.x

Mehta, U. M., Thirthalli, J., Naveen Kumar, C., Mahadevaiah, M., Rao, K., Subbakrishna, D. K., Gangadhar, B. N., & Keshavan, M. S. (2011). Validation of Social Cognition Rating Tools in Indian Setting (SOCRATIS): A new test-battery to assess social cognition. *Asian Journal of Psychiatry*, *4*(3), 203–209. https://doi.org/10.1016/j.ajp.2011.05.014

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Peña, E. D. (2007). Lost in Translation: Methodological Considerations in Cross-Cultural Research. *Child Development*, *78*(4), 1255–1264. https://www.jstor.org/stable/4620701

Prein, J. C., Kalinke, S., Haun, D. B. M., & Bohn, M. (2023). TANGO: A reliable, open-source, browser-based task to assess individual differences in gaze understanding in 3 to 5-year-old children and adults. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02159-5

Prein, J. C., Maurits, L., Werwach, A., Haun, D. B. M., & Bohn, M. (2024). *Variation in gaze following across the life span: A process-level perspective*. PsyArXiv. https://doi.org/10.31234/osf.io/dy73a

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*, *29*(1), 44–54. https://doi.org/10.3758/s13423-021-01948-3

R Core Team. (2024). *R: A language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing.

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

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *The Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Selcuk, B., Gonultas, S., & Ekerim-Akbulut, M. (2023). Development and use of theory of mind in social and cultural context. *Child Development Perspectives*, *17*(1), 39–45. https://doi.org/10.1111/cdep.12473

Shepherd, S. (2010). Following Gaze: Gaze-Following Behavior as a Window into Social Cognition. *Frontiers in Integrative Neuroscience*, *4*(5). https://doi.org/10.3389/fnint.2010.00005

Stengelin, R., Hepach, R., & Haun, D. B. M. (2020). Cultural variation in young children's social motivation for peer collaboration and its relation to the ontogeny of Theory of Mind. *PLOS ONE*, *15*(11), e0242071. https://doi.org/10.1371/journal.pone.0242071

Tang, Y., Gonzalez, M. R., & Deák, G. O. (2024). The slow emergence of gaze- and point-following: A longitudinal study of infants from 4 to 12 months. *Developmental Science*, *27*(3), e13457. https://doi.org/10.1111/desc.13457

Taumoepeau, M., Sadeghi, S., & Nobilo, A. (2019). Cross-cultural differences in children's theory of mind in Iran and New Zealand: The role of caregiver mental state talk. *Cognitive Development*, *51*, 32–45. https://doi.org/10.1016/j.cogdev.2019.05.004

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

Waschl, N., & Chen, M. (2022). Cross-Cultural Considerations for Adapting Valid Psychoeducational Assessments. In O. S. Tan, K. K. Poon, B. A. O'Brien, & A. Rifkin-Graboi (Eds.), *Early Childhood Development and Education in Singapore* (pp. 113–140). Springer. https://doi.org/10.1007/978-981-16-7405-1_7

Appendix

Selbstständigkeitserklärung

Julia Christin Prein
[Straße Hausnummer]
[PLZ Ort]
[Telefon]
[Email]

Hiermit erkläre ich, dass ich mich noch keiner Doktorprüfung unterzogen oder mich um Zulassung zu einer solchen beworben habe.

Ich versichere, dass die Dissertation [TODO: Titel Dissertation] in der gegenwärtigen oder einer anderen Fassung noch keiner anderen Hochschule zur Begutachtung vorgelegen hat.

Ich versichere an Eides statt, dass ich die eingereichte Dissertation [TODO: Titel Dissertation] selbstständig und ohne zulässige fremde Hilfe verfasst habe. Anderer als der von mir angegebenen Hilfsmittel und Schriften habe ich mich nicht bedient. Alle wörtlich oder sinngemäß anderen Schriften entnommenen Stellen habe ich kenntlich gemacht. Über die strafrechtlichen Folgen gemäß § 156 Strafgesetzbuch wurde ich in Kenntnis gesetzt.

[Ort], [Datum]   [Unterschrift]