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

Julia Christin Prein (ORCID: 0000-0002-3154-6167)[1,2], 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], 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], Umay Sen (ORCID: 0000-0001-9488-0851)[13], Katie Slocombe (ORCID: 0000-0002-7310-1887)[11], Robert Z. Sparks (ORCID: 0000-0001-7545-0522)[6], Roman Stengelin (ORCID: 0000-0003-2212-4613)[2,10], Jahnavi Sunderarajan[2], Kirsten Sutherland[2], Florence Tusiime[3], Wilson Vieira (ORCID: 0009-0001-9400-6328)[2], Zhen Zhang (ORCID: 0000-0001-9300-0920)[12], Yufei Zong (ORCID: 0009-0000-5012-0244)[12], Daniel B. M. Haun (ORCID: 0000-0002-3262-645X)[2,+], and & Manuel Bohn (ORCID: 0000-0001-6006-1348)[1,2,+]

[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
[13] Department of Psychology
Developmental Psychology
Uppsala University
[+] joint last author

Author Note

The authors made the following contributions. Julia Christin Prein (ORCID: 0000-0002-3154-6167): Conceptualization, Methodology, Software, Formal Analysis, Resources, Writing - Original Draft Preparation, Writing - Review & Editing; Florian M. Bednarski (ORCID: 0000-0003-4384-4791): Resources, Writing - Review & Editing; Ardain Dzabatou: Resources, Writing - Review & Editing; Michael C. Frank (ORCID: 0000-0002-7551-4378): Resources, Writing - Review & Editing; Annette M. E. Henderson (ORCID: 0000-0003-4384-4791): Resources, Writing - Review & Editing; Josefine Kalbitz: Resources, Writing - Review & Editing; Patricia Kanngiesser (ORCID:0000-0003-1068-3725): Resources, Writing - Review & Editing; Dilara Keşşafoğlu (ORCID: 0000-0002-7356-0733): Resources, Writing - Review & Editing; Bahar Köymen (ORCID: 0000-0001-5126-8240): Resources, Writing - Review & Editing; Maira V. Manrique-Hernandez: Resources, Writing - Review & Editing; Shirley Magazi (ORCID: 0009-0006-0479-9800): Resources, Writing - Review & Editing; Lizbeth Mújica-Manrique: Resources, Writing - Review & Editing; Julia Ohlendorf: Resources, Writing - Review & Editing; Damilola Olaoba: Resources, Writing - Review & Editing; Wesley R. Pieters (ORCID:0000-0002-6152-249X): Resources, Writing - Review & Editing; Sarah Pope-Caldwell: Resources, Writing - Review & Editing; Umay Sen (ORCID: 0000-0001-9488-0851): Resources, Writing - Review & Editing; Katie Slocombe (ORCID: 0000-0002-7310-1887): Resources, Writing - Review & Editing; Robert Z. Sparks (ORCID: 0000-0001-7545-0522): Resources, Writing - Review & Editing; Roman Stengelin (ORCID: 0000-0003-2212-4613): Resources, Writing - Review & Editing; Jahnavi Sunderarajan: Resources, Writing - Review & Editing; Kirsten Sutherland: Resources, Writing - Review & Editing; Florence Tusiime: Resources, Writing - Review & Editing; Wilson Vieira (ORCID: 0009-0001-9400-6328): Resources, Writing - Review & Editing; Zhen Zhang (ORCID: 0000-0001-9300-0920): Resources, Writing - Review & Editing; Yufei Zong (ORCID: 0009-0000-5012-0244): Resources, Writing - Review & Editing; Daniel B. M. Haun (ORCID: 0000-0002-3262-645X): Funding acquisition,

Writing - Review & Editing; Manuel Bohn (ORCID: 0000-0001-6006-1348): Conceptualization, Methodology, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Julia Christin Prein (ORCID: 0000-0002-3154-6167), Universitätsallee 1, 21335 Lüneburg, Germany. E-mail: julia.prein@leuphana.de

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: 5457

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) communities (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 universal and variable aspects of the human cognitive system cannot be drawn from data collected in single communities (Krys et al., 2024).

One reason for the lack of diversity in psychological studies is the scarcity of suitable methods to collect comparable psychological data in different communities (Bourdage et al., 2023). This shortage of suitable measures is even more prevalent in developmental psychology. In this paper, we describe the construction and psychometric evaluation of a measure of basic social cognition (gaze following) in children as a concrete example for how to overcome this problem.

Studies investigating variation between communities and/or individuals need to ensure that the captured variation is systematic and not just random noise: measures need to be reliable and valid. This, of course, applies to all fields of psychology. Given the topic of this paper, we will focus on measures of social cognition for children. Studies on social cognition 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 look at cross-cultural social cognition tasks (Bourdage et al., 2023; Hajdúk et al., 2020; Waschl & Chen, 2022). Thus, it is already 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 a pragmatic approach. The TANGO-CC was first implemented in one context, in our case Leipzig, Germany (Prein et al., 2023). Here, 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 (*i.e.*, where the participant thinks the target is) and the target's real 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

Table 1
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
Swahili	Bantu	Uganda
Turkish	Turkic	Türkiye
Yaka	Bantu	Rep. Congo

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.

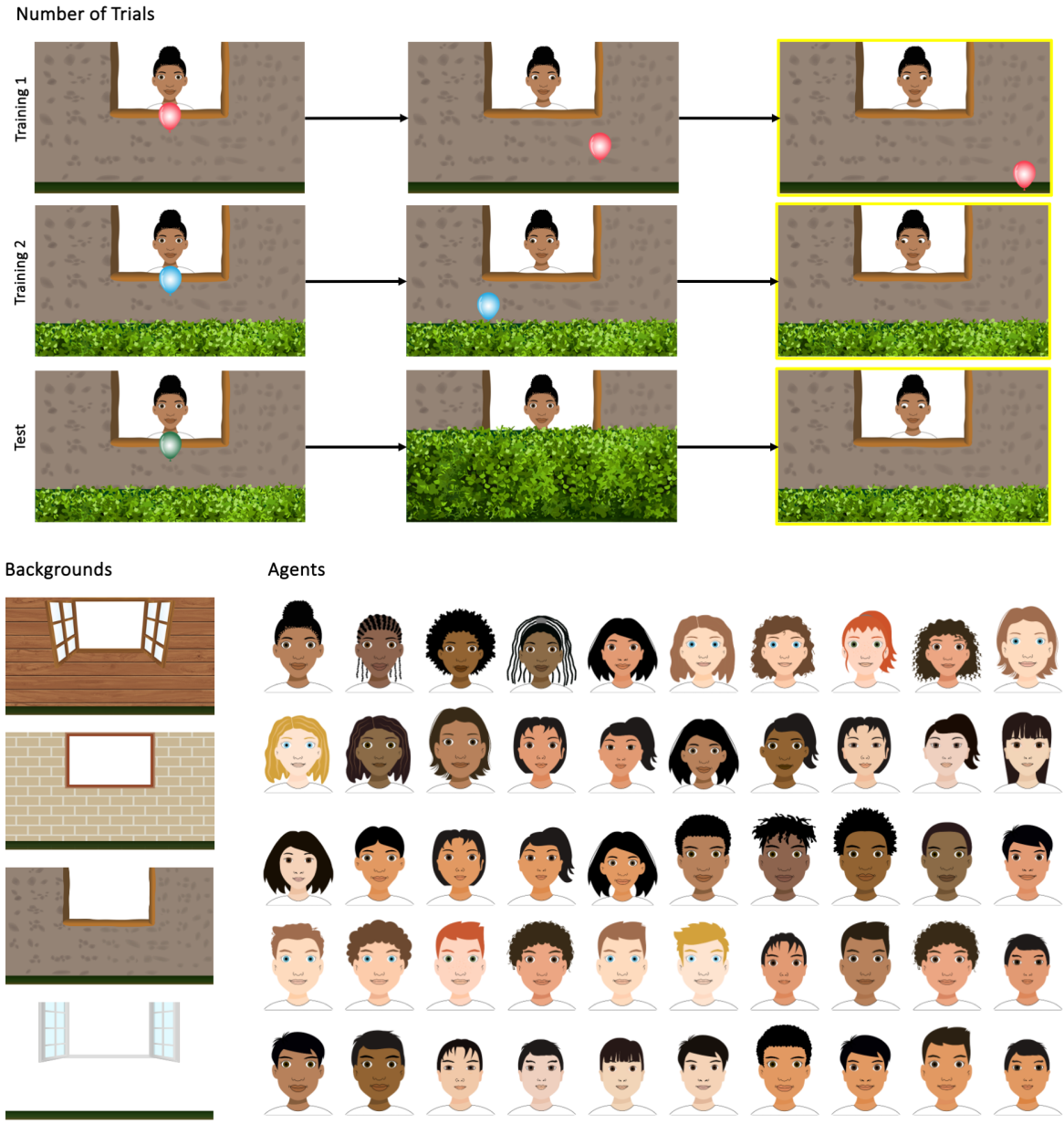

Figure 1

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

First, 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 (measured in 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 4.4.0 (2024-04-24) (R Core Team, 2024). GLMMs were fitted 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 fixed-effect 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 (random effects of trial type within community: minimum estimate for training 1 = -2.87; 95%CrI [-3.11; -2.60]; minimum estimate for training 2 = -1.27; 95%CrI [-1.51;

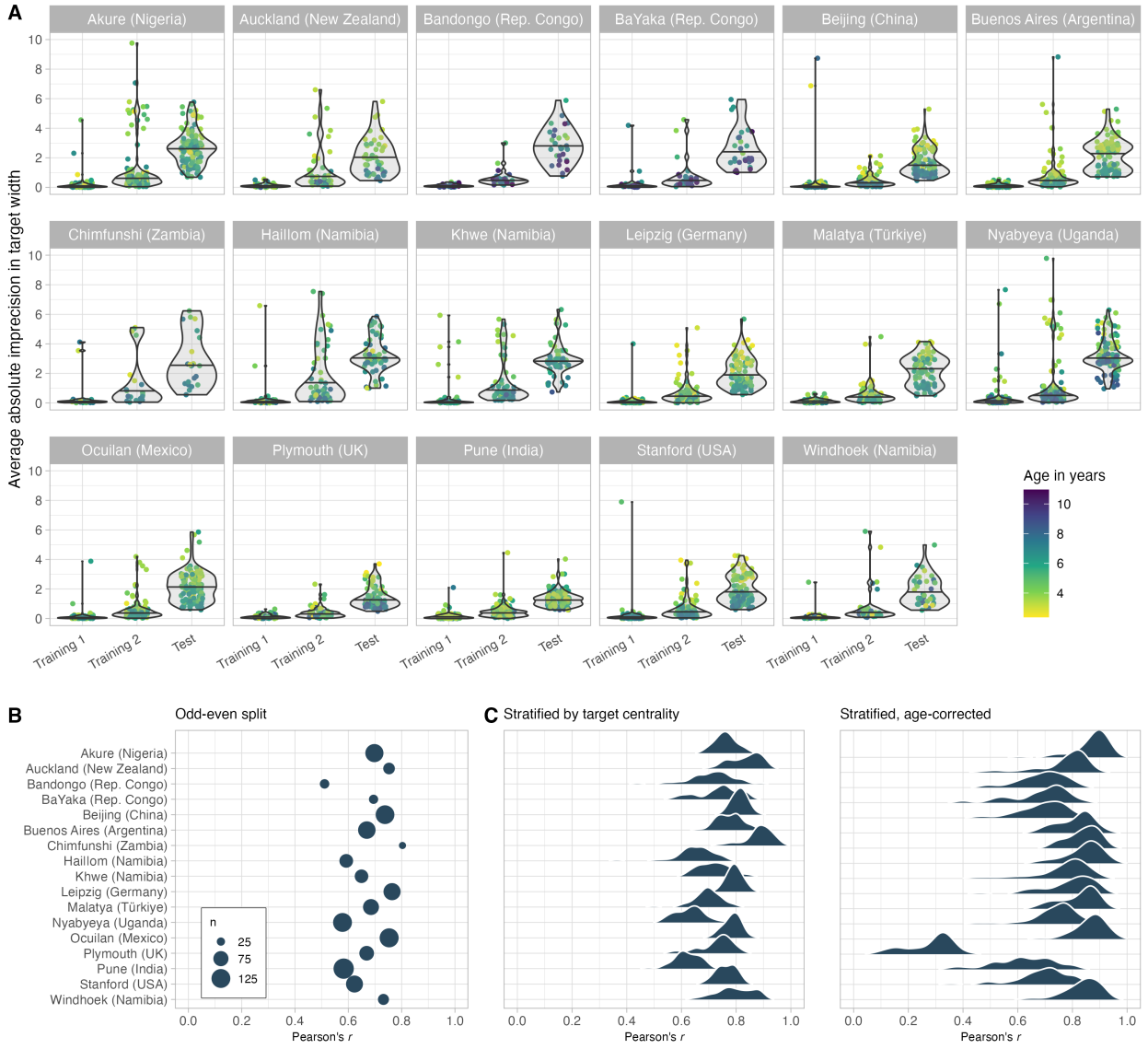

Figure 2

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.*

-0.98]). 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 substantial 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 fitted 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 indicate good internal consistency.[1] The results are comparable to the 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 better in the training than the test trials, and within-community variation greatly exceeded between-community variation. The task showed satisfactory to high reliability across all

[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.

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 captured 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 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