frontiers | Frontiers in Developmental Psychology

TYPE Original Research
PUBLISHED 25 September 2024
DOI 10.3389/fdpys.2024.1445406

OPEN ACCESS

EDITED BY
Deniz Tahiroglu,
Bogaziçi University, Türkiye

REVIEWED BY
Diane Poulin-Dubois,
Concordia University, Canada
Patricia Ann Prelock,
University of Vermont, United States

*CORRESPONDENCE
Franziska Baumeister
franziska.baumeister@unifr.ch

RECEIVED 07 June 2024
ACCEPTED 16 August 2024
PUBLISHED 25 September 2024

CITATION
Baumeister F, Wolfer P, Sahbaz S, Rudelli N, Capallera M, Daum MM, Samson AC, Corrigan G, Naigles L and Durrleman S (2024) Measuring Theory of Mind: a preliminary analysis of a novel linguistically simple and tablet-based measure for children. *Front. Dev. Psychol.* 2:1445406. doi: 10.3389/fdpys.2024.1445406

Measuring Theory of Mind: a preliminary analysis of a novel linguistically simple and tablet-based measure for children

Franziska Baumeister[1]*, Pauline Wolfer[1], Sümeyra Sahbaz[2], Nicola Rudelli[3], Marine Capallera[4], Moritz M. Daum[5], Andrea C. Samson[6,7], Grace Corrigan[8], Letitia Naigles[8] and Stephanie Durrleman[1]

[1]ABCCD Lab, Faculty of Science and Medicine, University of Fribourg, Fribourg, Switzerland, [2]Institute for Research, Development and Evaluation, Bern University of Teacher Education, Bern, Switzerland, [3]Department of Education and Learning/University of Teacher Education, Competence Centre for School, Social and Educational Needs (BESS), University of Applied Sciences and Arts of Southern Switzerland, Locarno, Switzerland, [4]HumanTech Institute, University of Applied Sciences and Arts of Western Switzerland, HES-SO, Fribourg, Switzerland, [5]Developmental Psychology: Infancy and Childhood, Department of Psychology, University of Zurich, Zurich, Switzerland, [6]Faculty of Psychology, UniDistance Suisse, Brig, Switzerland, [7]Department of Special Education, University of Fribourg, Fribourg, Switzerland, [8]Child Language Lab, Psychological Sciences, University of Connecticut, Storrs, CT, United States

This study introduces a novel linguistically simple, tablet-based, behavioral Theory of Mind (ToM) measure, designed for neurotypical (NT) and autistic children aged 4–10 years. A synthesis of five comprehensive reviews of existing ToM measures revealed significant gaps in their designs; the weaknesses include a mismatch between the operational and conceptual definition of ToM, high verbal demands in most measures, materials that are minimally interesting for children, and often a lack of psychometric evaluations. These findings call into question the suitability of most of the currently available ToM measures used in children, both with and without developmental disorders, such as children with autism spectrum disorder (ASD). For example, the assessment of ToM in children with ASD may require reduced reliance on complex language or social interaction that can be part of the diagnostic criteria of the condition. This newly designed ToM measure, developed in line with the "Standards for Educational and Psychological Testing" of the American Educational Research Association, is linguistically simple, tablet-based, suitable for children with ASD, and is available in English, German, French, Italian, and Spanish. With a sample of 234 participants, including 152 NT children and 82 children with ASD between 4 and 10 years of age, the new ToM measure's psychometric properties were preliminarily evaluated. Descriptive statistics, measures of internal consistency, inter-item correlation, and validity checks were conducted in both groups. Further inspections of the measure's scale- and item-level characteristics were conducted with the help of exploratory factor analyses (EFA), and item response theory (IRT) within the NT children's group. These preliminary evaluations suggest that the newly developed ToM measure possesses good psychometric properties and is both accessible and engaging for children. Further investigation with a larger group of participants is necessary to reinforce these initial results.

This will allow item- and scale-level assessments within a wider range of autistic children. For this purpose, the task will be made freely available to the scientific community.

KEYWORDS

Theory of Mind, measurement, children, autism, validation, psychometric properties, tablet-based

1 Introduction

Theory of Mind (ToM) is a vital cognitive ability, involving understanding and inferring mental states such as desires, beliefs, and intentions, grasping that these may differ between oneself and others, and predicting ensuing behaviors (Wimmer and Perner, 1983).

Besides, ToM is an important predictor of reading comprehension and inevitable for social interaction (Jacobs and Paris, 1987; Slaughter et al., 2015). As recent reviews reveal inconsistencies in existing ToM measures (Ziatabar Ahmadi et al., 2015; Beaudoin et al., 2020; Quesque and Rossetti, 2020; Osterhaus and Bosacki, 2022; Fu et al., 2023), assessing ToM with currently existing measures presents challenges. These highlighted gaps explain the need for new measures like our newly created ToM measure discussed in the present paper.

1.1 ToM as a developmental ability

Researchers agree that ToM develops through a sequential process in early and middle childhood, identifiable via a series of different tasks. For instance, Wellman and Liu (2004) examined the age at which neurotypical (NT) children displayed proficiency in various domains of ToM, which included understanding other people's desires, beliefs, knowledge, and emotions. The examination of a series of subtasks subsequently led to the development of a ToM scale consisting of seven items that measure the developmental progression of ToM in children: (1) a "Diverse Desires" task, testing the ability to judge that two individuals may have different desires about the same objects, which is generally considered to start developing at the age of two; (2) a "Diverse Beliefs" task, testing the ability to judge that two individuals may have different beliefs about the same objects, considered to develop around the age of three; (3) a "Knowledge Access" task, testing the ability to judge the knowledge of another individual who does not share the participant's knowledge, a competency that is considered to develop around 3–4 years of age; a series of "False Belief" tests which develop between 4–5 years of age, namely (4) a "Contents False Belief" task, testing the ability to judge another individual's false belief about the content of a container, such as a Band-aid box; (5) an "Explicit False Belief" task, testing the ability to predict a subsequent behavior of another individual who has a false belief; (6) a "Belief Emotion" task, testing the ability to judge how another individual will feel, based on a false belief; and finally (7) a "Real-Apparent Emotion" task, testing an individual's ability to judge that a person can feel something but display a different emotion, a skill which develops between 5–6 years of age (Wellman and Liu, 2004). So-called "higher-order reasoning" is considered to develop later in middle childhood, around the age of seven or eight. The latter is for example measured with Second-order False Belief tasks that require understanding that someone may hold a false belief about someone else's belief (Perner and Wimmer, 1985).

1.2 ToM in autistic children

Difficulties in ToM were once considered to present the main "deficit" in individuals diagnosed with autism spectrum disorder (ASD)[1] (Baron-Cohen et al., 1985; Baron-Cohen, 1997). ASD is a neurodevelopmental disorder characterized by difficulties in initiating and sustaining social interaction and social communication, repetitive behavior, as well as hypersensitivity to sensory stimuli (World Health Organization, 2019). Additionally, due to the heterogeneous nature of the condition, autistic children may encounter language development delays across multiple linguistic levels, such as in the morphosyntactic or syntactic domains (Naigles, 2021; Schaeffer et al., 2023; Silleresi, 2023). Children with ASD were considered to have a ToM "deficit". The normative perspective leading ToM to be considered as the "core" deficiency and "lack" in ASD resulted in the notion of "mindblindness" (Baron-Cohen, 1997). In contrast, contemporary studies recognize that children on the spectrum may attain ToM insights with differing rates and magnitudes (Peterson and Wellman, 2019; Marocchini, 2023).

1.3 Importance of ToM

ToM is linked to various other cognitive skills, such as inhibitory control and working memory (Joseph and Tager-Flusberg, 2004), and linguistic abilities, such as syntactic comprehension (De Villiers and Pyers, 2002). Moreover, ToM is associated with and is a significant predictor of reading (Jacobs and Paris, 1987), defined as the ability to extract meaning from written text. Specifically, ToM is of paramount importance in analyzing and monitoring the mental states of characters involved in narratives, enabling the reader to deduce information regarding these mental states even without explicit statements about them.

1 We will use both person-first and identity-first language interchangeably when referring to individuals diagnosed with ASD, to acknowledge the diverse preferences within the autistic community (Vivanti, 2020; Bottema-Beutel et al., 2021; Buijsman et al., 2023).

Difficulty with this type of inference is particularly evident among autistic individuals (Dore et al., 2018). Furthermore, ToM is considered crucial for social interaction because it allows individuals to understand and respond to the emotions and beliefs of others, which in turn facilitates peer collaborations and seamless social interactions and prevents miscommunications (Slaughter et al., 2015). Considering the importance of understanding ToM in autistic and NT children to ultimately guide clinical practices and provide individualized support for children with ToM difficulties, it is crucial for researchers to utilize ToM measures designed explicitly for the populations under investigation.

1.4 Findings from systematic reviews of ToM measures

To provide a comprehensive overview and evaluation of the existing ToM measures, we present the findings from five recent systematic reviews (Ziatabar Ahmadi et al., 2015; Beaudoin et al., 2020; Quesque and Rossetti, 2020; Osterhaus and Bosacki, 2022; Fu et al., 2023). These systematic reviews summarize the current state of the existing ToM measures and reveal their limitations for accurately and consistently assessing ToM in children aged 4–10 years with and without ASD. Furthermore, these reviews have examined the suitability of the existing ToM measures for assessing ToM in children and have evaluated various aspects, such as presentation modes, psychometric properties, alignment with ToM definitions, and classification of tasks within the ToM framework. A review by Ziatabar Ahmadi et al. (2015) focused on ToM tests available in English to assist researchers and clinicians in selecting an appropriate tool to evaluate social cognition. Beaudoin et al. (2020) examined 220 measures specifically designed for children between zero and 5 years old. This review focused on the modes of presentation, scoring options, and psychometric properties of these measures. As a result, the authors developed a framework called "Abilities in Theory of Mind Space" (ATOMS), which identified ToM subcategories. Fu et al. (2023) reviewed studies in which 127 different ToM measures were utilized. Their analysis focused on the measures' constructs, modes of presentation and response, theories predicting ToM development, and psychometric properties. Quesque and Rossetti (2020) conducted a review that assessed ToM measures based on their level of correspondence to the underlying ToM definition. Osterhaus and Bosacki (2022) created a systematic review of advanced ToM measures, focusing on ToM definitions and task classification.

As Beaudoin et al. (2020, p. 1) pointed out, "identifying appropriate assessment tools for young children remains challenging". First, an important claim across the reviews is that there is often a mismatch between the *conceptual* and the *operational* definition of ToM: On one hand, the conceptual definition of ToM can encompass a variety of abilities, such as understanding desires, beliefs or emotions. On the other hand, a series of operational translations of this concept have resulted in tasks that measure some components of ToM, such as diverse desires, first-order false beliefs, second-order false beliefs or emotion recognition. Additionally, the broad nature of ToM makes it challenging to assess all aspects of its developmental and multidimensional nature, in part due to time constraints. Consequently, the tasks employed to measure "Theory of Mind" also exhibit substantial variation or overlap, despite the authors' distinct conceptualizations of ToM. This raises concerns about the validity of the measures' score interpretations, as authors cannot ascertain whether they are truly measuring ToM or something else. Therefore, when operationalizing ToM, examining its alignment with the conceptual definition is crucial. Second, the length of ToM measures is hugely variable, with most existing measures containing one item and several others containing a large number of items. Although measures with few items are quicker to administer, they raise concerns about the measure's reliability (Ziatabar Ahmadi et al., 2015). Third, another concern highlighted in the reviews is that the existing ToM measures place too great a burden on cognitive resources, which may increase difficulty for participants, particularly those with ASD (Fu et al., 2023). For instance, the reviews claim that most ToM measures require a high level of verbal comprehension to understand the tasks, which can be challenging for children with limited verbal abilities. Fu et al. (2023) and Georgopoulos et al. (2022) thus suggest favoring tasks with visual aids and multiple-choice responses to support children. Fourth, concerning the involvement of human interaction during the assessment, Fu et al. (2023) underscore that many ToM measures rely on direct interaction with the child, often using read-aloud stories or picture scenarios. A claim is therefore made to use more varied formats, including videos or audio recordings. These formats control for inter-tester variability while providing a more engaging and less intimidating setting for children, particularly those with ASD. Fifth, the review authors raise concerns about the psychometric robustness of the ToM measures, as most ToM measures have not been subjected to examinations of validity and reliability. In this respect, Ziatabar Ahmadi et al. (2015) emphasize the need for a systematic approach to creating and assessing measures, focusing on internal consistency, inter-rater reliability, test-retest reliability, and criterion validity. Internal consistency refers to the degree of correlation between different items within a measurement tool and indicates how reliably these items measure the same underlying construct; inter-rater reliability describes the degree of consistency between two or more raters' scores of verbally assessed responses; test-retest reliability indicates the stability of test scores over time when the same test is administered to the same group of participants at two separate time points; criterion validity indicates the extent to which scores obtained from a new measurement tool correlate with an established measurement tool evaluating the same construct. The use of measures meeting these psychometric criteria would thus be crucial for valid and cross-laboratory comparable interpretations of ToM performances. Fu et al. (2023) furthermore suggest using item response theory (IRT). IRT is an item-level assessment, allowing to establish a link between the properties of the items on a measure, the participants' so-called "trait levels" on this measure, and the underlying trait being measured (Morizot et al., 2007).

The use of measures meeting these psychometric criteria would thus be crucial for valid and cross-laboratory comparable interpretations of ToM performances. Among the tools discussed in the five reviews, the "Theory of Mind Task Battery" by Hutchins

et al. (2008), Hutchins and Prelock (2016) stands out as a promising direct behavioral ToM assessment, testing a range of ToM abilities with evaluated psychometric properties. However, this tool includes only one item per tested construct. Indirect assessments through parental questionnaires, which include several items per construct, provide an alternative. Four questionnaire measures were identified in the reviews, including the "Children's Social Understanding Scale" (CSUS) by Tahiroglu et al. (2014) and the Theory of Mind Inventory-2 (ToMI-2) questionnaire by Hutchins and Prelock (2016) and Hutchins et al. (2012). The CSUS targets children between 3 and 7 years and assesses understanding of belief, knowledge, desire, intention, perception and emotion through 42 items. In contrast, the ToMI-2, consisting of 60 items, focuses on early, basic and advanced ToM abilities. Both questionnaires can be answered by caregivers regarding their children's ToM abilities in daily life. However, because parental evaluations may not offer precise insights into these abilities, though, researchers such as Beaudoin et al. (2020) argue that it is important to integrate behavioral laboratory assessments alongside such questionnaire evaluations.

1.5 Standards for Educational and Psychological Testing

Given the need for a new behavioral ToM measure that addresses the aforementioned claims, the newly created behavioral ToM measure is designed to adhere to the main claims of the manual "Standards for Educational and Psychological Testing" (American Educational Research Association et al., 2014) to ground our research within recognized guidelines. The manual provides a foundational framework for test development in education and psychology. It underscores the importance of validity and reliability while also considering accessibility, fairness in testing, and the beneficial impact of technology on testing. Following the guidelines, as well as our intended use case of assessing ToM in NT and autistic children between 4 and 10 years of age, the following criteria were important: Firstly, the concept (Theory of Mind) under investigation needs to be well-defined and the purpose of the measure and the interpretation of its scores must be presented explicitly. Secondly, the measure has to be valid. That is, it needs to accurately measure the construct it is intended to measure: in our case, ToM. Validity pertains to interpreting test scores rather than the test itself, emphasizing the necessity of a context-specific validation process for each test application. Thirdly, the tool needs to be reliable. That is, it must produce consistent results across various items and ability levels. Fourthly, the tool should adhere to the guidelines listed concerning fairness in testing, the use of technology, and test development. In terms of accessibility and fairness of testing, any elements not related to the specific construct being measured need to be minimized, as these may impede a participant's comprehension of instructions or their ability to respond accurately. This includes adjusting factors such as language complexity to suit participant proficiency levels. This is especially crucial when testing populations with potential linguistic difficulties, to prevent disadvantaging individuals who have difficulty in understanding complex linguistic constructions. Concerning the criterion of a universal design, planning assessments with the target population's diverse needs in mind is key to ensuring the test's appropriateness and utility. This means making the test useful for all subgroups within the intended population, such as considering varying language abilities and age ranges. Regarding standardizing the test environment, providing a consistent test environment for all participants is essential. This encompasses clear and concise instructions, specified time limits, suitable testing spaces, qualified test administrators, and uniform technology to hinder inconsistencies and unfairness. Furthermore, developing a psychological or educational test involves ensuring its effectiveness, validity, and user-friendliness. Key considerations include the purpose, its target audience, the test length, and the respondent's burden. Balancing precision with practicality is essential, as longer tests may yield more precise results but also increase participant fatigue. Efficient stop criteria can help manage test duration effectively, preventing disengagement (American Educational Research Association et al., 2014).

1.6 This paper

The goal of the present research paper is to present a novel, linguistically simple, tablet-based, behavioral ToM measure. It was created as part of a PhD project within the "Autism, Bilingualism, Cognitive, and Communicative Development" (ABCCD) project at the University of Fribourg, Switzerland, and is therefore referred to as the "ABCCD ToM measure". During the PhD project on the impact of bilingualism on ToM in children with ASD, the need for an appropriate tool was stated for the target population (autistic and NT children between the ages of 4 and 10). This paper shows the preliminarily psychometric tests in a pilot study involving NT adults and in a main study with NT children and autistic children. The analyses include descriptive measures of mean performance on the test items of the task, tests of internal consistency that examine the degree of agreement between the items of included sub-measures of ToM, and construct validity tests that indicate the extent to which scores in sub-domains correlate with subscores in the ToMI-2, an existing ToM measure (Hutchins et al., 2012; Hutchins and Prelock, 2016). These psychometric evaluations are followed by exploratory factor analysis (EFA), and item response theory analysis (IRT) to inspect the measure's item- and scale characteristics further (Bean and Bowen, 2021). The presentation of the results forms the basis for discussing the utility of the ABCCD ToM measure in assessing specific components of ToM in scientific and clinical settings. The study procedure was approved by the Swiss Ethics Research Committee and the Institutional Review Board of the University of Connecticut, USA.

2 Materials and methods

2.1 The ABCCD ToM measure

2.1.1 The creation process

2.1.1.1 Inspiration from existing ToM measures

The ABCCD ToM measure draws inspiration from several existing ToM measures: the verbal ToM scale by Wellman and Liu

(2004) consists, as presented earlier, of seven subtasks with one item each to assess the understanding of diverse desires, diverse beliefs, diverse knowledge, two types of false beliefs, and emotions. The same types of items were also used in the "low-verbal" adaptation by Burnel et al. (2018) who integrated one item per assessment of the understanding of diverse desires, diverse beliefs, knowledge access, content false belief, and explicit false belief. The different item types of both measures and the linguistically simple nature of Burnel's measure have been established as important criteria for our new behavioral ABCCD ToM measure. However, Wellman et al.' and Burnel et al.' scales have two significant shortcomings for our target population: first, both scales consist of only one item per sub-ability tested; second, since the abilities tested are considered to be acquired in NT children before the age of seven, these measures are not suitable for assessing NT children much older than 7 or 8 years. Therefore, inspiration was taken additionally from other ToM measures: The ToM measure by Forgeot d'Arc and Ramus (2011) consists of a series of First-order False Belief tasks displayed with the help of video clips without complex narrations. Furthermore, the recently created ToM toolkit by Marinis et al. (2023) developed Forgeot d'Arc and Ramus' measure further. It integrated Second-order False Belief items, which are considered to be performed correctly in NT children around the age of 7 or 8 (Miller, 2009). Marinis et al.' measure also contains clips with linguistically simple narrations which presents a suitable assessment of higher-order ToM in children with ASD. However, the Second-order False Belief items in this toolkit present one important shortcoming: The structure of the videoclips together with the response choices displayed at the end allows a correct answer to second-order items not only by applying second-order reasoning but also first-order reasoning (for more details, please see the Supplementary material). We, therefore, integrated into our newly created behavioral ABCCD ToM measure the strengths of the measures we drew inspiration from while paying attention to the shortcomings concerning psychometric properties and participants' burden of existing ToM measures overall. Ideally, the measure would have consisted of a series of cognitive and affective aspects of ToM, as tested in Wellman and Liu's ToM scale (2004), while adding a higher-order task and reducing the language complexity, or in the "ToM Task Battery" by Hutchins and Prelock (2016), while adding more items per construct and reducing the language complexity. However, such a measure would have resulted in a very long assessment, potentially leading to participant fatigue and lack of concentration, so that we had to select a limited number of sub-abilities to integrate. Therefore, our final ABCCD ToM measure consists of three test blocks: "Diverse Desires" (Block 1), "First-order False Beliefs" (Block 2), and "Second-order False Beliefs" (Block 3).

Addressing the claim in previous ToM measure reviews of an often-observed mismatch between the conceptual and operational definition of ToM, the conceptual definition of ToM in this work specifically englobes the ability to understand that other individuals may have different perspectives, specifically desires and beliefs, and to attribute these perspectives to others, whether false or correct, in scenarios of increasing complexity. The operational definition and interpretation of test scores consequently includes assessment of the understanding of diverse desires (Block 1), first-order false beliefs (Block 2), and second-order false beliefs (Block 3).

The three blocks were selected because they represent key milestones in the sequential development of ToM in NT children, with each ability building upon the previous one. The assessment of the understanding of diverse desires was included because it marks an early developmental milestone in ToM, typically emerging around the age of 2–3 years in NT children (Wellman and Liu, 2004). This construct is crucial for identifying difficulties in early ToM abilities, which are known to occur in autistic children (Broekhof et al., 2015). First-order false belief understanding was chosen because it assesses whether children have reached the critical developmental stage of recognizing that others can hold beliefs different from their own, and that these subjective beliefs may not align with objective reality. This ability generally develops around 4–5 years in NT children (Wellman and Liu, 2004). Given its frequent use in studies involving children with ASD, including in the ABCCD ToM measure, facilitates comparability with previous research. Second-order false belief understanding was included as it represents a more advanced level of ToM, involving the understanding that others can have beliefs about another person's belief. This level of recursive belief reasoning typically develops around the age of 7–8 years in NT children (Perner and Wimmer, 1985). By including this construct, we aim to assess more advanced ToM abilities in older children. For NT children it was therefore expected that children of different ages are able to pass the different blocks at the aforementioned ages, while we expected similar or later acquisition in the autistic group.

During the creation process of the three blocks, in cooperation with programmers, various discussions and reviews by research experts and members of the autistic community informed the design of the tasks. This was followed by individual tests with few adults and children. The final version of the ABCCD ToM measure was ultimately piloted with 40 NT adults. The analyses showed that adults, as expected, perform at ceiling on all items within all constructs.[2]

2.1.1.2 Tablet-based assessment

The new ABCCD ToM measure was created in a gamified way as an application to be run on tablets, viewed by the child without the necessity of a direct interaction with the experimenter. We collaborated with a team of programmers to develop the visually animated items for each block using the game engine "Unity" (Haas, 2014, version: 2020.3.48f1). All tasks are presented in a 3-dimensional gamified environment to provide a visually engaging experience for the participant. The virtual circus setting, accompanied by the presence of an animated character named Gabi, aims to create an immersive and enjoyable atmosphere. The visual setup and colors were kept minimal to focus only on essential story elements and avoid distractions and sensory overload (World Health Organization, 2019). Such a tablet-based assessment can support reducing potential validity problems mentioned earlier. Firstly, by presenting the same instructions and items to each participant, the quality of the presentation of the tasks is controlled

2 For a detailed overview of the results of the pilot study with 40 NT adults, please see the Supplementary material.

in advance. Secondly, assessments of cognitive abilities with the help of a tablet have proven effective in autistic children (Alzrayer et al., 2014; Alhajeri et al., 2017). By including appealing comic-like characters, the task was tailored to be attractive to children. Thirdly, the potential burden on participants regarding emotional effort and sensory overload can be minimized since the need for direct interaction with the test administrator is not necessary thanks to the tablet-based assessment.

2.1.1.3 Linguistic complexity of narrations

Given that children with ASD may exhibit developmental language delays (Rapin and Dunn, 2003), care was taken to ensure that all crucial information to solve the tasks is presented visually and accompanied by the simplest verbalization possible. The linguistic complexity of the narrations was kept to a minimum on the lexical (Tager-Flusberg and Sullivan, 1994), morphosyntactic (Roberts et al., 2004), and syntactic levels (Durrleman et al., 2016) to avoid any linguistic features that may pose difficulties for children with ASD (Burnel et al., 2018). In this respect, only short main clauses in the present tense were used. The verbal narrations in the languages of testing (i.e., English, French, Italian, German, Spanish) were created adhering to the standard for cross-linguistic translations, by using translation and back-translation. The narrations were initially crafted in English, then translated and adjusted by two researchers in their respective native languages before being checked again with the English version. Critical parts of the narrations, including test questions, were carefully reviewed by the team of researchers to finally reach a consensus. Native speakers of English, French, Italian, German, and Spanish recorded the narrations in a soundproof room. All audio recordings were edited using "Audacity" (Audacity Team, 2014).

The task materials, including scripts, videos, and a manual on how to create adaptations in other languages, can be accessed on the OSF repository (Section 5): https://osf.io/pg2an/?view_only=ce9836db48d7477db52f79db7187995b.

2.1.2 Structure of the ABCCD ToM measure

The ABCCD ToM measure consists of three blocks and each block consists of 2 practice items, 2 control items and 4 test items.

2.1.2.1 Block 1—Diverse Desires

Block 1 (Diverse Desires) tests the participant's understanding that two individuals, specifically the participant and another individual ("Gabi") may "have different desires about [. . .] objects" (Wellman and Liu, 2004, p. 531) or activities. The block consists of eight items, including two practice items, two control items, and four test items, and is divided into four timeframes (T_1 to T_4, for an example see: https://osf.io/56dca), visible in Table 1. In T_1, the circus director presents three different objects or three different activities to both the participant and to Gabi. For example, at one point the director says: "We have balls, ropes, and hula hoops." In T_2, the director inquires about the participant's preference among the objects or activities by naming them and pointing to them, accompanied by a visual highlight. The participants indicate their preference by clicking on one of the objects or activities. In T_3, Gabi is asked the same question to determine his preference. Gabi responds by verbally naming his desired object or activity, which is also visually highlighted. Finally, in T_4, Gabi is prompted to take his preferred object or the object for his preferred activity, and the participant is asked: "What will Gabi take?" The participant needs to respond by clicking on Gabi's desired object or on the object for Gabi's desired activity.

The items are presented in two conditions. The first is the "Diverse Desires" condition, which is the test condition, where Gabi decides on an object or an activity that differs from the participant's choice. In this condition, the participant must differentiate between his or her and Gabi's desire. The second is the "Equal Desires" condition, serving as the control and practice condition, where Gabi wants the same object or activity as the participant. This condition ensures that the participant comprehends the task (practice items) and controls whether they are correctly following the scenarios (control items). At the end of each item, three response choices are presented. In the "Diverse Desires" condition, the correct answer is Gabi's choice, the incorrect answer is the participant's choice, and the oddball answer is the third option. A correct answer results in a score of "1", while the incorrect and the oddball answer yield a score of "0". A correct answer to the test items is interpreted as indicating that participants can distinguish between their own and someone else's desires. In contrast, an incorrect answer to the test items is interpreted as attributing their desire to someone else, and an oddball answer is interpreted as not understanding the task. A scale score can be obtained by summing up the scores on the four test items, leading to a "Diverse Desires scale score" between 0 and 4. In case both control items from the "Equal Desires" condition are answered incorrectly, the "Diverse Desires scale score" will be set to 0 to rule out the possibility that the participant did not understand the task or did not pay attention.

2.1.2.2 Block 2—First-order False Beliefs

Block 2 (First-order False Beliefs) tests the participants' understanding that other individuals can hold different beliefs about reality. This ability is examined by asking "how someone will search [or what someone will first do], given that person's mistaken belief about reality" (Wellman and Liu, 2004, p. 531).

Each scenario in Block 2 is divided into four timeframes (T_1 to T_4, for an example see: https://osf.io/qdta8), visible in Table 2. In T_1, two circus performers are introduced (referred to as "Circus Performer 1" and "Circus Performer 2"). For instance, in one test item, a clown and an acrobat are playing "Hide and Seek" in a room where three objects are visible: a chair, a bed, and a sofa. The clown starts counting while the acrobat hides behind the bed. During this time, the clown sees the acrobat moving behind the bed. In T_2, a change is undertaken by Circus Performer 1 that is visible (in practice and control items) or invisible (in test items) to Circus Performer 2. For example, the acrobat moves behind the sofa without being seen by the clown. In T_3, Circus Performer 2 is ready to take action. In the given example, the clown turns around and wants to search for the acrobat. In T_4, three possible endings are presented through short video clips, depicting what will logically happen first ("Now what will happen first?"). In the provided example, the endings may involve the clown looking either behind the chair, bed, or sofa. The participant needs to respond by picking one of the three response choices.

The items are presented in two conditions: "False Belief" items and "True Belief" items. The False Belief items represent the test

TABLE 1 Overview of the structure of the scenarios in Block 1 (example: test item 2).

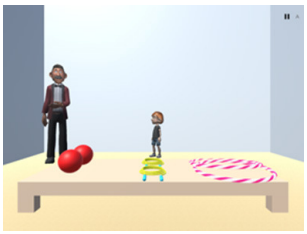
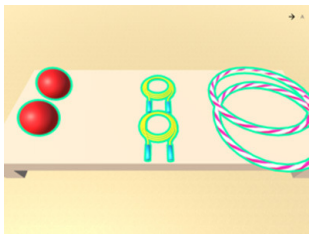
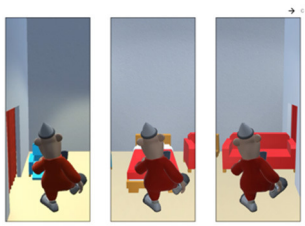
T_1 Set up of scenario	T_2 Desire participant	T_3 Desire Gabi	T_4 Target question on Gabi's desire
			
"We have balls, ropes, and hula hoops."	"What do you want: A ball, a rope, or a hula hoop?"	"And you, Gabi, what do you want: a ball, a rope, or a hula hoop?"	"What will Gabi take?"

TABLE 2 Overview structure of the scenarios in Block 2 (example: test item 2).

T_1 Set up of scenario	T_2 Change	T_3 End	T_4 Target question
			
The clown and the acrobat start playing "Hide and Seek"; the acrobat moves behind the bed and is observed by the clown.	The acrobat moves, without being seen by the clown, behind the sofa.	The clown turns around and starts searching for the acrobat.	"Now what will happen first?"

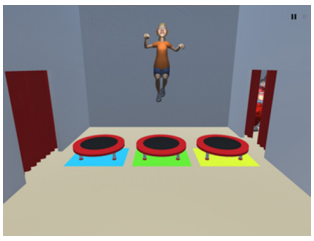
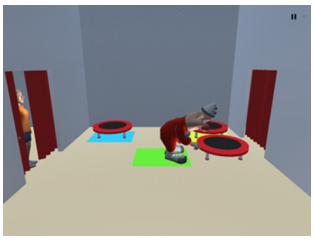
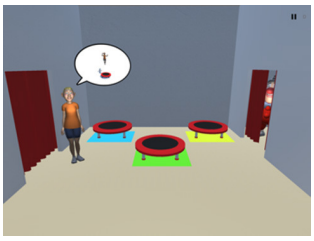
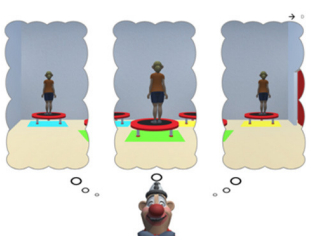
condition (Dennett, 1978) in which Circus Performer 2 is unaware of the change made by Circus Performer 1. In this condition, participants are required to distinguish between their perspective, which aligns with reality, and Circus Performer 2's perspective, which is based on a false belief about reality. This condition assesses the participants' ability to attribute a false belief to Circus Performer 2, such as predicting what will happen first based on Circus Performer 1's false belief. The True Belief items serve as the control and practice conditions where Circus Performer 2 observes the change made by Circus Performer 1. The participant and Circus Performer 2 share the same perspective and belief regarding what will happen first. This condition ensures that the participants understand the task (practice items) and controls whether they are correctly following the scenarios (control items).

At the end of each item, three response choices are presented. For a question following a False Belief (test) item, the accurate response is the one that considers the perspective of Circus Performer 2 who has a mistaken belief due to not witnessing a change that occurred, such as the clown in the mentioned example. An inaccurate response is one that considers the participant's perspective that aligns with reality, while an oddball answer is the third response choice. A correct answer will result in a score of "1", an incorrect answer in a score of "0," and an oddball answer in a score of "0". A correct answer to the test items is interpreted as indicating that the participant can attribute a false belief to someone else. In contrast, an incorrect answer to the test items is interpreted as attributing the participant's perspective to someone else, and an oddball answer is interpreted as not understanding the task. A scale score can be obtained by summing up the scores on the four test items, leading to a "First-order False Beliefs scale score" between 0 and 4. In case both control items from the True Belief condition are answered incorrectly, the "First-order False Beliefs scale score" will be set to 0 to rule out the possibility that the participant did not understand the task or did not pay attention.

2.1.2.3 Block 3—Second-order False Beliefs

Block 3 (Second-order False Beliefs) tests the understanding that another person can hold a false belief about the belief of a third person. Specifically, it examines the ability to predict what a person with a false belief about a third person's belief will do (Perner and Wimmer, 1985). Each item in Block 3 is divided into four timeframes (T_1 to T_4, for an example see: https://osf.io/wpemd), illustrated in Table 3. In T_1, the scene is set with Circus Performer 2 entering the stage and performing an action, while being observed by Circus Performer 1. For instance, in one test item, an acrobat tries out three different trampolines, two of which are faulty and one that works well. The clown observes this situation through a doorway. In T_2, Circus Performer 1 enters the stage and changes the scene. Circus Performer 2 witnesses this change, but Circus Performer 1 is (in test items only) unaware that Circus Performer 2 is watching. In the given example, the clown changes the position of the good and one of the faulty trampolines while being seen

TABLE 3 Overview structure of the scenarios in Block 3 (example: test item 2).

T_1 Set-up of scenario	T_2 Change	T_3 End	T_4 Target question
			
The acrobat tries out the three trampolines, realizing that only the one in the middle works well.	The clown enters and changes the good one with one of the faulty trampolines, while being observed by the acrobat. However, the clown is not aware that the acrobat is watching.	The acrobat comes back.	"Now what does the clown think will happen first?"

by the acrobat. However, the clown is unaware of the fact that the acrobat was watching the change of the trampolines. In T_3, Circus Performer 2 re-enters the room, ready to continue the action initiated in T_1. In the example item, the acrobat now returns willing to jump, and the clown's head appears behind the curtain. In T_4, three alternative endings are presented through short video clips, depicting three different thoughts of Circus Performer 1 about what will happen now ("Now what does the clown think will happen first?"). In the provided example, one response choice shows the clown thinking that the acrobat will go onto the faulty trampoline that the clown had changed with the good trampoline. Another shows the clown thinking that the acrobat will go onto the good trampoline that the clown had changed with one of the faulty trampolines. The third response choice presents the clown thinking that the acrobat will go onto the other faulty trampoline that had not been touched.

The items are presented in two conditions: "False Belief" items and "True Belief" items. The False Belief items represent the test condition (Dennett, 1978) in which Circus Performer 1 does not know that Circus Performer 2 has observed the change. In this condition, participants are required to distinguish between their perspective, which aligns with reality and Circus Performer 2's perspective, and Circus Performer 1's perspective, which is based on a false belief about Circus Performer 2's perspective. This condition, therefore, assesses the participants' ability to attribute a false belief to Circus Performer 1, such as predicting what Circus Performer 1 thinks will happen first. The True Belief items serve as the control and practice conditions where Circus Performer 1 knows that Circus Performer 2 observed the change. The participant and Circus Performer 1 thus share the same perspective and belief regarding what Circus Performer 1 thinks will happen first. This condition ensures that the participants get familiar with the task (practice items) and controls whether they are correctly following the scenarios (control items).

At the end of each item, three response choices are presented. For a question following a False Belief (test) item, the accurate response is the one that considers the perspective of Circus Performer 1, who is holding a mistaken belief about Circus Performer 2's perspective. An inaccurate response is the one that considers the participant's perspective that aligns with reality and with Circus Performer 2's perspective, while an oddball answer is the third response choice. A correct answer will result in a score of "1", an incorrect answer in a score of "0" and an oddball answer in a score of "0" as well. A correct answer to the test items is interpreted as indicating that the participant can attribute a belief to someone else with a false belief about another person's perspective. In contrast, an incorrect score to the test items is interpreted as attributing the participant's own perspective to someone else and an oddball answer is interpreted as not understanding the task. A scale score can be obtained by summing up the scores on the four test items, leading to a "Second-order False Beliefs scale score" between 0 and 4. In case both control items from the True Belief condition are answered incorrectly, the "Second-order False Beliefs scale score" will be set to zero to rule out the possibility that the participant did not understand the task or did not pay attention.

2.1.2.4 Stop criteria between blocks

Stop criteria were implemented between the three blocks. If a participant did not obtain a scale score above 1 in a block, the subsequent block was not shown to the participant; this criterion was implemented in accordance with the guidelines presented in Section 1 to minimize the participants' burden.

2.2 Participants

The study participants comprised 234 individuals, as shown in Table 4: 152 NT children, and 82 children with ASD.

2.2.1 NT children

One hundred fifty-two NT children between 4 and 10 (mean age: 7;2) participated in the study in Switzerland, Germany, France, Canada, and the USA. Eighty-three identified as female, 68 as male, and one as gender-diverse. All participants included in the final data set reported no history of language or cognitive delays or impairments and no history of a diagnosis of ASD. NT children were recruited through flyers, advertisement emails through primary school contacts, participant databases of previous projects, and Facebook. The parents of all participants gave written informed consent; the children received a gift card for participating in the project (35CHF/35€/60CAD/35USD).

TABLE 4 Participant overview.

	Neurotypical children ($N = 152$)	Autistic children ($N = 82$)
Age		
Mean in months (SD)	86.3 (24.5)	95.8 (24.5)
Mean in years; months (range)	7;2 (4;0–10;11)	7;12 (4;2–10;11)
Gender		
Diverse (*N*, in %)	1 (0.7%)	0 (0.0%)
Female (*N*, in %)	83 (54.6%)	12 (14.6%)
Male (*N*, in %)	68 (44.7%)	70 (85.4%)
Parental educational level		
Mean (SD; range)	4.63 (0.72; 1–5)	3.94 (1.25; 1–5)
Non-verbal IQ (RPM, z-score)		
Mean (SD; range)	98.5 (12.7; 66–133)	95.4 (14.2; 66–133)
Receptive vocabulary (PPVT, z-score)		
Mean (SD; range)	0.45 (1.13; −2.33–2.73)	−0.85 (1.68; −4.0–2.67)

RPM, Raven's Progressive Matrices (Raven et al., 2018); PPVT, Peabody Picture Vocabulary Test (Dunn et al., 1993, 2016; Dunn and Dunn, 2007; Lenhard et al., 2015).

2.2.2 Autistic children

Eighty-two children with ASD between 4 and 10 (mean age: 7;12) participated in the study in Switzerland, Germany, France, Canada, and the USA. Twelve identified as female, 70 as male. All participants included in the final data set had an official diagnosis of ASD, assessed with either the Autism Diagnostic Observation Schedule-2nd Edition (ADOS-2; Lord et al., 2003) or another standardized ASD diagnosis tool, such as the Autism Diagnostic Interview-Revised (Lord et al., 1994). Autistic children were recruited through flyers, advertisement emails through autism associations, psychologists, speech and language therapists, official recruitment platforms ("BuildClinical", USA), and Facebook. The parents of all participants gave written informed consent; the children received a gift card for participating in the project (35CHF/35€/60CAD/35USD).

2.3 Testing procedure

The children were enrolled in the study between February 2023 and April 2024. The participants' caregivers were first asked to complete online questionnaires which took ~60 min, using "Gorilla Experiment Builder" (Anwyl-Irvine et al., 2020): one background questionnaire on the children's personal history, the Q-BEx (De Cat et al., 2022) on the children's language experiences, the "Social Communication Questionnaire (SCQ)" (Rutter et al., 2003), and the SWAN (Swanson et al., 2012). Based on these questionnaires, a decision about including and excluding children in the study was taken. Autistic children were able to enroll if they had an official diagnosis of ASD, provided by a clinician; NT children were able to enroll if they had no diagnosis of ASD or any other neurodevelopmental disorder. The parents were also asked to fill in the ToMI-2 (Hutchins and Prelock, 2016). The ToMI-2 consists of 60 statements; the parents are asked to judge on a scale with scores between zero and 20 to what extent the given statement holds for their child. The questionnaire's items can be subdivided into three different subscales that children are considered to develop in sequential order: an "early" subscale (including items testing abilities that are achieved in infancy and toddlerhood), a "basic" subscale" (including metarepresentation skills, developed around and after 4 years of age), and an "advanced" subscale (including items assessing more complex forms of recursion and social judgment) (Hutchins et al., 2012).

The ABCCD ToM measure was administered in person at the children's homes ($N = 201$) or schools ($N = 33$) on iPad through the application in the children's most proficient language. All participants completed Block 1 (Diverse Desires), Block 2 (First-order False Beliefs), and Block 3 (Second-order False Beliefs), unless they reached the stop criterion so that no subsequent block was shown. The entire ToM measure lasted up to 30 min, including breaks. The ToM measure was followed by other tasks part of the study protocol, including a short form of the RPM (Raven et al., 2018) to control for nonverbal cognitive abilities, as well as the PPVT (French: Dunn et al., 1993, Italian: Dunn et al., 2016; English: Dunn and Dunn, 2007; German: Lenhard et al., 2015), to control for receptive vocabulary.

2.4 Analysis plan

All analyses were conducted using the statistical computing environment R (R Core Team, 2020; version: 4.3.3).

2.4.1 Preliminary analyses

We conducted preliminary analyses and assessed with the help of logistic mixed effects regression models (Bates et al., 2015) whether the place of testing (school or home) and language of testing (English, French, German, Italian, or Spanish) influenced ToM performance. For the assessment of the place of testing, *testing place* was included as a fixed effect, and *participant* and *item* as random effects, while controlling for *age*, *language proficiency*, and *working memory*. For the assessment of the language of testing, a very similar model was built with the sum-coded effect of *language of testing* as fixed effect.

2.4.2 Mean performance

We examined the mean performances separately for both NT and autistic children across the four test items in each of the blocks (Diverse Desires, First-order False Beliefs, Second-order False Beliefs).

2.4.3 Internal consistency

We checked the internal consistency of each block separately for each group. We calculated the Kuder-Richardson 20 (Kuder and Richardson, 1937), with the arbitrary cutoff of KR-20 > 0.70, which is widely considered as presenting acceptable to good

internal consistency (e.g., Ntumi et al., 2023), as well as inter-item correlations. For the latter, scores below 0.2 may indicate that items measure different constructs, whereas scores above 0.7 may indicate redundancy (Röschel et al., 2021).

2.4.4 Validity argumentation

We conducted a series of analyses to evaluate the validity of the ABCCD ToM measure's scores. Since no highly similar ToM measure to the ABCCD ToM measure existed, we examined the correlations between subscale scores on the ABCCD ToM measure and the subscale scores of the ToMI-2 (Hutchins et al., 2012). The ToMI-2 is known for its strong psychometric properties, and includes measures of constructs of increasing complexity, which are similar to those integrated into the ABCCD ToM measure. The ToMI-2 comprises "early", "basic", and "advanced" subscales while the ABCCD ToM measure includes diverse desires understanding, first-order false-beliefs, and second-order false-beliefs. In contrast, the CSUS includes measures of diverse desires and first-order false beliefs but does not assess second-order false beliefs (Tahiroglu et al., 2014). Furthermore, the ToMI-2 was designed for an age group comparable to that in our study (Hutchins et al., 2012), whereas the CSUS was intended for a younger age group (Tahiroglu et al., 2014).

Specifically, we calculated correlations between the "Diverse Desires scale score" (scale score of Block 1 of the ABCCD ToM measure) and the "Early subscale score" of the ToMI-2, the "First-order False Beliefs scale score" (scale score of Block 2 of the ABCCD ToM measure) and the "Basic subscale score" of the ToMI-2, as well as the "Second-order False Beliefs subscale score" (scale score of Block 3 of the ABCCD ToM measure) and the "Advanced subscale score" of the ToMI-2. We anticipated only low to moderate correlations due to the ToMI-2's broader range of targeted ToM abilities and its reliance on parental judgement rather than direct assessment of the child's abilities.

Additionally, we created three logistic mixed effect regression models to assess whether the ABCCD ToM measure is sensitive to (a) age-related changes (Hutchins and Prelock, 2016), (b) linguistic abilities, and (c) cognitive abilities that develop in parallel with ToM, such as language development and working memory (Carlson et al., 2004; Milligan et al., 2007). In the first model (a), we included *age* as a fixed effect and *participant* and *item* as random effects; in the second model (b), we replaced *age* with *language proficiency*. In the third model (c), we replaced *age* with *working memory*.

To determine whether the ABCCD ToM measure can distinguish between NT and autistic children, given that autistic children are known to present ToM difficulties (e.g., Tager-Flusberg, 2007), we created an additional model that included *diagnostic group* as a fixed effect while controlling for *age*, *IQ*, *language proficiency*, and *working memory*, and added *participant* and *item* as random effects.

To further evaluate the ABCCD ToM measure's scale- and item-level psychometric properties, we utilized exploratory factor analyses (EFA), using the "lavaan" package (Rosseel, 2012; version: 0.6-17) and item response theory (IRT), using the "mirt" package (Chalmers, 2012; version: 1.41). EFA and IRT were however only conducted for NT children since the limited sample size of the autistic group did not allow meaningful analyses (Morizot et al., 2007).

2.4.5 Exploratory factor analysis (EFA)

For a scale-level evaluation of the psychometric properties of the ABCCD ToM measure, we conducted an EFA. Thus, we evaluated the factor structure of the ABCCD ToM measure's subscales' scores in NT children and examined the unidimensionality assumption required for IRT. Because the three blocks consist of binary items, we evaluated the factor structure of each block with the Weighted Least Squares Mean and Variance (WLSMV)/Diagonally Weighted Least Squares (DWLS) estimator and the promax oblique rotation method. Model fit was assessed using the Chi-square statistic, its p-value, and Comparative Fit Index (CFI). Since we fit models with only two degrees of freedom while having a small sample size, model fit assessment with the help of Root Mean Square Error of Approximation (RMSEA) is not recommended (Kenny et al., 2015; Shi et al., 2022).

2.4.6 Item response theory (IRT)

To provide an item-level evaluation of the psychometric properties of the ABCCD ToM measure, we used IRT. IRT is a framework that relies on models and assumes the existence of a latent trait influenced by a person's responses and the parameters of the items. This framework enables the estimation of the item parameters and the trait levels of participants simultaneously (Reise et al., 2005). Its models establish a connection between the characteristics of the items in a measurement, the ability levels of individuals who respond to these items, and the latent trait being assessed, in our case understanding of diverse desires, first-order false beliefs, and second-order false beliefs. IRT assumes that each participant has a position on the latent trait (also called θ) that influences the probability that a participant will select a particular item response category. The mathematical relationship between each item and θ, estimated by fitting IRT models, is characterized by a slope and a location parameter. The slope or "discrimination" parameter provides information about the extent to which items can distinguish between individuals with different levels of the underlying latent trait θ. Values exceeding 1.34 suggest a strong discriminatory effect (Baker, 2011). Conversely, the location or item "difficulty" parameters reveal the trait level at which participants have a probability of 0.5 of selecting a higher response option, thereby shedding light on the item's capacity to capture various trait levels. There can be more than one location parameter in the case of items with more than two response choices. As we are presenting a task that only contains dichotomous variables, we will only speak about "one" difficulty parameter. For sample sizes between 100 and 200 (Morizot et al., 2007), so-called "Rasch modeling" or "1-Parameter Logistic modeling" is suggested. It is characterized, as opposed to "2-Parameter Logistic models", by estimating only the difficulty parameter of each item, with a fixed discrimination parameter set to 1, implying that all items are assumed to differentiate between individuals at the same rate. We report the results of the 1-Parameter Logistic (1PL) models, using a full information marginal maximum likelihood

fitting function as a function of the fixed weights of the predictors for the models. To run IRT, we followed the approach suggested by Morizot et al. (2007) who propose starting with a check of the assumptions, before fitting the model. Therefore, after testing the assumption of unidimensionality utilizing EFA for each block (Diverse Desires, First-order False Beliefs, and Second-order False Beliefs), we fitted 1PL IRT models to examine the measurement precision of each item across the ability continuum. The models for each subtask were subsequently assessed with the help of "Item difficulty" parameters, which can also be illustrated in an "Item Characteristic Curve" (ICC). The ICC demonstrates the latent trait level on the x-axis, ranging from low to high. The y-axis represents the probability of a correct response to the individual test items, ranging from 0 to 1. A higher probability indicates a greater likelihood that the item will be answered correctly, given the respondents' latent trait levels. The shape of the curve should be S-shaped, indicating that as the latent trait level increases, the probability of a correct response also increases. The item difficulties are indicated by the point on the latent trait axis where the probability of a correct response is 0.5. Items with curves that shift more toward the right indicate greater difficulty, as a higher level of the latent trait is required to have a 50% chance of a correct response. The ICC enables the evaluation of a second assumption, monotonicity, which suggests that as the trait levels increase, the likelihood of a correct response also increases. Furthermore, within the framework of IRT, the item information is determined by the latent trait and is derived from the participant's responses to the items. The "Item Information Function" (IIF) serves as a metric for statistical information provided by an item, similar to the reliability of measurement, which indicates the accuracy of an item across the latent trait continuum (Reise et al., 2005). Consequently, items with high information at a specific latent trait level offer precise estimations of person parameters within that level of the latent trait continuum (Baker, 2002).

3 Results

The results section presents the findings from the examination of the psychometric properties of the ABCCD ToM measure for NT and autistic children, following the analysis plan described above. The data, the commented R-Markdown source code and the HTML output of the analyses can be found on OSF (Section 1): https://osf.io/pg2an/?view_only=ce9836db48d7477db52f79db7187995b.

3.1 Preliminary analyses

Among the 234 children, 33 were tested in school, whereas 201 were tested in their homes. Analyses using logistic mixed effects modeling, including *testing place*, *age*, *language proficiency*, and *working memory* as fixed effects and *item* and *participant* as random effects, indicated that the testing place did not significantly predict ToM performance (Estimate = −0.278, SE = 0.557, z-value = −0.499, $p = 0.618$).

Children completed the ABCCD ToM measure either in English, French, German, Italian or Spanish. Although the narrations accompanying the items were created in a very careful manner to be identical in each language version, we verified with the help of logistic mixed effect regression modeling whether there were differences across the language versions. Analyses showed that none of the language versions compared to the "grand mean" showed a significant difference (English: Estimate = −0.613, SE = 0.733, z-value = −0.835, $p = 0.403$, French: Estimate = 0.482, SE = 0.608, z-value = −0.792, $p = 0.428$, German: Estimate = 0.989, SE = 0.627, z-value = 1.579, $p = 0.114$, Italian: Estimate = −0.632, SE = 0.676, z-value = −0.936, $p = 0.349$, Spanish: Estimate = 0.738, SE = 1.325, z-value = 0.557, $p = 0.578$).

TABLE 5 Percentages of correct responses to individual items in Blocks 1, 2, and 3 in NT children.

Block 1		Block 2		Block 3	
Diverse Desires		First-order False Beliefs		Second-order False Beliefs	
$N = 152$		$N = 145$		$N = 96$	
Test item 1	96.1%	Test item 1	51.7%	Test item 1	68.8%
Test item 2	92.1%	Test item 2	55.2%	Test item 2	66.7%
Test item 3	94.7%	Test item 3	57.2%	Test item 3	57.3%
Test item 4	93.4%	Test item 4	66.2%	Test item 4	51.0%

3.2 Neurotypical children

3.2.1 Mean performances

In NT children, as shown in Tables 5, 6, performance in Block 1 was close to ceiling, with weaker performances, as expected, in Blocks 2 and 3, therefore consequently leading to a substantial number of participants to whom subsequent blocks were not presented due to the implemented stop criterion; theses cases are presented in Table 6 in gray.

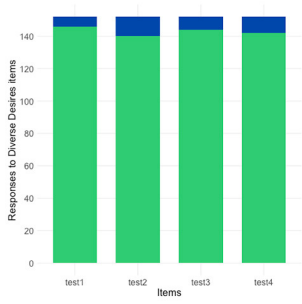
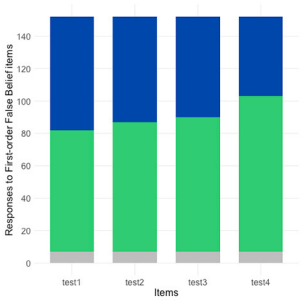
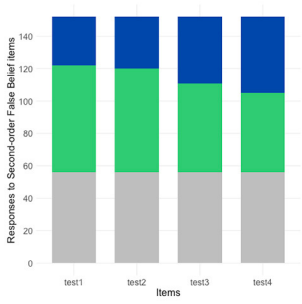
To get an idea about when children can be considered to "pass" a block, that is, answer correctly to at least 2 out of 4 test items, the age ranges were inspected separately.

Figure 1 shows that Block 1 was "passed" by children at the age of 4 in over 95% of the cases, in children at the age of 5 in over 87% of the cases, and with higher percentage rates for children 6 years and older. Block 2 was passed by only 14% of children at the age of 4, 19% of children at the age of 5, 41% of children at the age of 6, and over 50% to 85% of children at the ages of 7 and older. Block 3 was passed by less than 20% of children younger than 6, 38% of children at 7, and over 72% of children older than 8. For a detailed overview of the percentages of correct responses to individual items, please see Appendix 3.

3.2.2 Internal consistency

The internal consistency of each block was measured with the Kuder-Richardson 20 score (Kuder and Richardson, 1937). In Block 1 (Diverse Desires), the KR-20 was 0.81, and in both Block 2 (First-order False Beliefs) and Block 3 (Second-order False Beliefs), the KR-20 was 0.83. The inter-item correlations within Blocks 1 and 2

TABLE 6 Types of responses to individual items in Blocks 1, 2, and 3 in NT children.

Block 1	Block 2	Block 3
Diverse Desires	First-order False Beliefs	Second-order False Beliefs
$N = 152$	$N = 145$	$N = 96$
		

Green: number of correct responses; blue: number of incorrect responses; gray: number of participants who did not run the block (because the previous block was not passed).

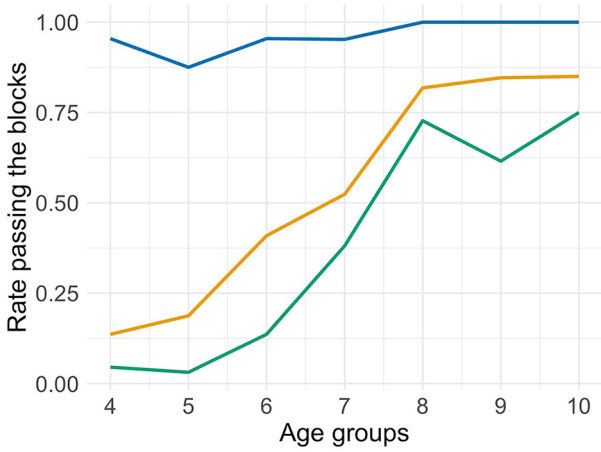


FIGURE 1
Rate of NT children at different ages passing Blocks 1, 2, and 3. Blue line: Block 1 (Diverse Desires), orange line: Block 2 (First-order False Beliefs), green line: Block 3 (Second-order False Beliefs).

were both 0.54 and in Block 3 0.55; thus, indicating no redundancy among the items within blocks.

3.2.3 Validity argumentation

The validity of the ABCCD measure scores was evaluated through several analyses. First, correlations between the ABCCD ToM subscales' scores and the subscales' scores of the ToMI-2 indicated varying degrees of association. Specifically, the correlation between the "Diverse Desires subscale scores" (Block 1) and the ToMI-2's "Early subscale scores" was weak and nonsignificant ($r = 0.12$). In contrast, the correlation between the "First-order False Beliefs subscale scores" (Block 2) and the ToMI-2's "Basic subscale scores" was significant ($r = 0.33$, $p < 0.05$). Similarly, the correlation between the "Second-order False Beliefs subscale scores" (Block 3) and the ToMI-2s "Advanced subscale scores" was significant ($r = 0.34$, $p < 0.05$).

Furthermore, the analysis revealed that (a) older children performed significantly better than younger children (Estimate = 0.088, SE = 0.011, z-value = 8.320, $p < 0.001$), (b) children with higher language proficiency outperformed those with lower proficiency (Estimate = 0.903, SE = 0.246, z-value = 3.671, $p < 0.001$), and (c) children with better working memory abilities performed better than those with weaker working memory abilities (Estimate = 1.123, SE = 0.172, z-value = 6.525, $p < 0.011$).

3.2.4 Exploratory factor analysis (EFA)

We conducted an EFA to examine the factor structure across Blocks 1, 2, and 3. The analyses utilized DWLS estimation with promax oblique rotation. In Block 1 (Diverse Desires), the chi-square statistic was 0.792 with 2 degrees of freedom, resulting in a p-value of 0.67, which indicates an excellent fit (CFI = 1). The first eigenvalue was significantly higher than subsequent values, suggesting a strong dominant factor that explained 83.8% of the variance. Item loadings ranged from 0.82 to 0.98, indicating high correlations with the dominant factor. The communalities varied from 0.67 to 0.97, supporting a high degree of variance explanation by the factor. In Block 2 (First-order False Beliefs), the chi-square statistic was 0.69 with 2 degrees of freedom, and a p-value of 0.71, reflecting a perfect model fit (CFI = 1). This block also featured a prominent dominant factor, explaining 77.2% of the variance. Item loadings were strong, between 0.79 and 0.99. Communalities ranged from 0.62 to 0.97. In Block 3 (Second-order False Beliefs), the chi-square statistic was 2.92 with 2 degrees of freedom, and a p-value of 0.23, indicating a good fit (CFI = 0.95). A single factor accounted for 79.5% of the total variance. Item loadings varied from 0.82 to 0.93, with all items showing a strong correlation with the factor. Communalities for this block were between 0.67 and 0.87. The EFA results thus provide robust evidence of a single-factor structure within each block.

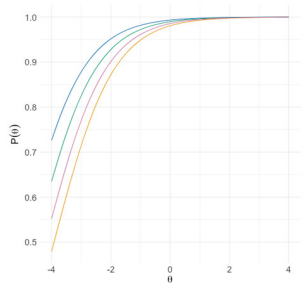
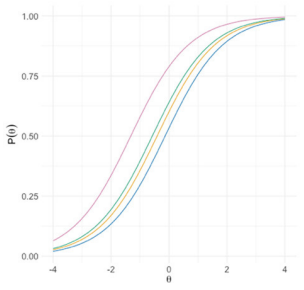
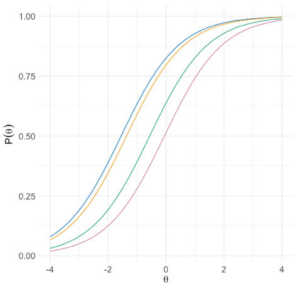
3.2.5 Item response theory (IRT)

After examining unidimensionality with the EFA and the local independence in the variable structure, we conducted IRT. The item difficulty parameters for the four items in each block are

TABLE 7 Item difficulty parameter estimators for all four test items within each block in NT children.

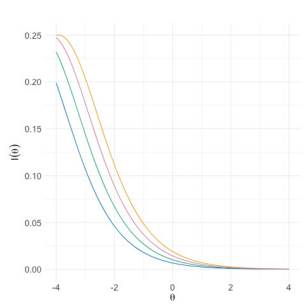
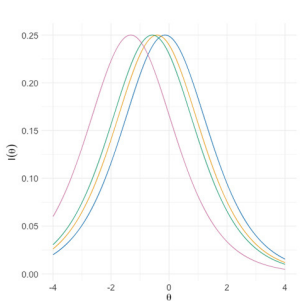
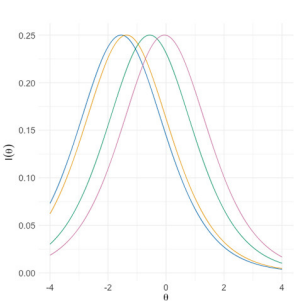
Block 1 Diverse Desires $N = 152$		Block 2 First-order False Beliefs $N = 145$		Block 3 Second-order False Beliefs $N = 96$	
Item	Difficulty	Item	Difficulty	Item	Difficulty
Test item 1	−4.98	Test item 1	−0.14	Test item 1	−1.55
Test item 2	−3.92	Test item 2	−0.41	Test item 2	−1.36
Test item 3	−4.55	Test item 3	−0.58	Test item 3	−0.56
Test item 4	−4.21	Test item 4	−1.32	Test item 4	−0.05

TABLE 8 Item characteristics curves for each block in NT children.

Block 1 Diverse Desires	Block 2 First-order False Beliefs	Block 3 Second-order False Beliefs
		

Blue line: test item 1 (within each block); orange line: test item 2 (within each block); green line: test item 3 (within each block); purple line: test item 4 (within each block).

TABLE 9 Item information function for each block in NT children.

Block 1 Diverse Desires	Block 2 First-order False Beliefs	Block 3 Second-order False Beliefs
		

Blue line: test item 1 (within each block), orange line: test item 2 (within each block), green line: test item 3 (within each block), lilac line: test item 4 (within each block).

TABLE 10 Percentages of correct responses to individual items in Blocks 1, 2, and 3 in autistic children.

Block 1 Diverse Desires $N = 82$		Block 2 First-order False Beliefs $N = 54$		Block 3 Second-order False Beliefs $N = 26$	
Test item 1	65.8%	Test item 1	46.3%	Test item 1	69.2%
Test item 2	74.4%	Test item 2	44.4%	Test item 2	69.3%
Test item 3	74.4%	Test item 3	51.9%	Test item 3	57.7%
Test item 4	69.5%	Test item 4	50.0%	Test item 4	57.7%

presented in Table 7. All discrimination parameters were set to 1 as we used Rasch modeling. The item difficulty values varied in Block 1 between −4.98 and −3.92, Block 2 between −1.32 and −0.14, and Block 3 between −1.55 and −0.05.

Table 8 provides the item information curves (ICC) for the four items within the three blocks. We see from the ICCs for all three blocks that the items present relatively small difficulty. Given, however, that only participants performing "well" on a previous block were also shown the subsequent block, we would have assumed lower performance of those participants on Blocks 2 and 3, and therefore higher item difficulties in these two blocks. That means, even if the latent trait levels best captured by the items seem to be below 0 for Blocks 2 and 3, we assume that the curves, in reality, would be shifted more toward the right on the x-axis.

Table 9 provides the item information functions (IIF) for the four items within the three blocks. The x-axis represents the latent trait level (Diverse Desires, First-order False Beliefs, Second-order False Beliefs), ranging from low to high levels. The y-axis represents the information an individual item provides about the latent trait. Higher information implies greater precision in estimating a respondent's trait level. We see that the four items within each block present slightly different item information and therefore capture well different trait levels. Some items seem to have similar item information, such as items 2 and 4 in Block 2 or items 1 and 2 in Block 3, that can indicate similar ability to capture a participant's trait level at a specific trait level.

3.3 Autistic children

3.3.1 Descriptive analyses

In autistic children, performance was descriptively slightly weaker in comparison to NT children. As shown in Tables 10, 11, all items seemed to be descriptively equally difficult when inspecting the participants' mean performance.

To get an idea about when children can be considered to "pass" a block, that is, answer correctly to at least 2 out of 4 test items, the age ranges were inspected separately. Figure 2 shows that Block 1 was "passed" by children at the age of 4 in over 57% of the cases, in children at the age of 5 in 50% of the cases, in about 62% at the ages of 6 and 7, and in more than 70% of the cases in older children. Block 2 was not passed by more than 10% of children until the age of 6, 18% of children at the age of 7, 20% at the age of 8, and around 40 to 55% at the ages of 9 and 10. Block 3 was not passed by more than 10% of children until the age of 9, 55% at the age of 9, and 35% at the age of 10. For a detailed overview of the percentages of correct responses to individual items, please see Appendix 3.

3.3.2 Internal consistency

In Block 1, the KR-20 was 0.9, in Block 2 0.8, and in Block 3 0.75, indicating good internal consistency. The inter-item correlation in Block 1 was 0.68, in Block 2 0.5, and in Block 3 0.44.

3.3.3 Validity argumentation

The correlation between the "Diverse Desires subscale scores" (Block 1) and the ToMI-2's "Early subscale scores" was weak but

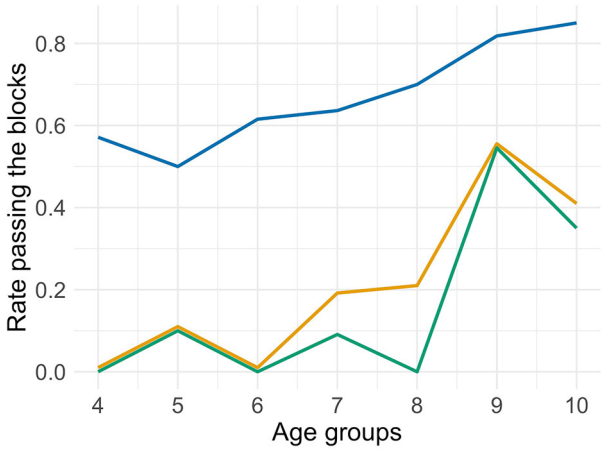
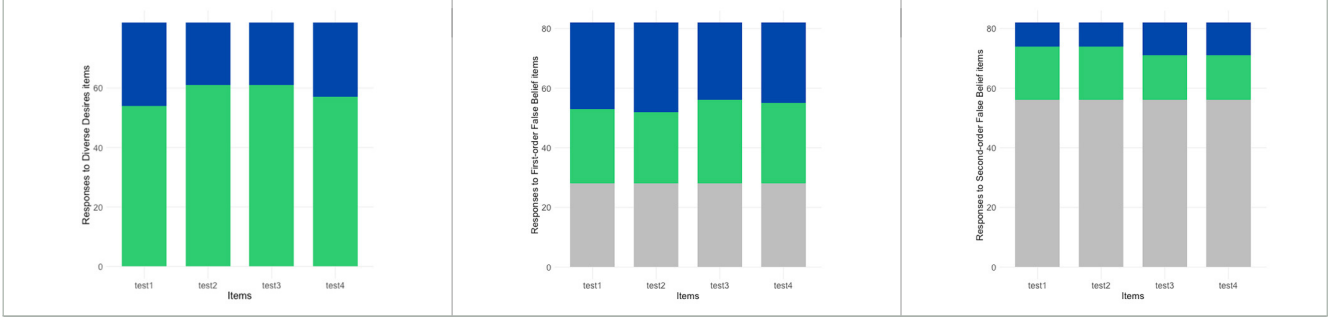


FIGURE 2
Rate of autistic children at different ages passing Blocks 1, 2, and 3. Blue line: Block 1 (Diverse Desires), orange line: Block 2 (First-order False Beliefs), green line: Block 3 (Second-order False Beliefs).

TABLE 11 Types of responses to individual items in Blocks 1, 2, and 3 in autistic children.

Block 1	Block 2	Block 3
Diverse Desires	First-order False Beliefs	Second-order False Beliefs
$N = 82$	$N = 54$	$N = 26$
		

Green: number of correct responses; blue: number of incorrect responses; gray: number of participants who did not run the block (because the previous block was not passed).

significant ($r = 0.22$, $p < 0.05$). The correlation between the "First-order False Beliefs subscale scores" (Block 2) and the ToMI-2's "Basic subscale score" was low to moderate but nonsignificant ($r = 0.23$, $p = 0.09$). The correlation between the "Second-order False Beliefs subscale scores" (Block 3) and the ToMI-2's "Advanced subscale score" was moderate but nonsignificant ($r = 0.33$, $p = 0.1$).

Furthermore, the analysis revealed that (a) *age* significantly predicted ToM performance (Estimate = 0.070, SE = 0.017, z-value = 4.157, $p < 0.001$), (b) *language proficiency* was a significant predictor (Estimate = 1.239, SE = 0.267, z-value = 4.633, $p < 0.001$) and (c) *working memory* also significantly predicted ToM performance (Estimate = 0.856, SE = 0.239, z-value = 3.586, $p < 0.001$).

When including both NT and autistic children in a model while controlling for *age*, *IQ*, *language proficiency* and *working memory*, the effect of *participant group* (NT vs. autistic children) was significant (Estimate = −1.727, SE = 0.516, z-value = −3.346, $p < 0.001$). This finding indicates that NT children performed significantly better than autistic children, suggesting that the ABCCD ToM measure effectively discriminates between these groups.

4 Discussion

4.1 Summary of the rationale for the creation process of a new Theory of Mind measure

This study introduced and evaluated the ABCCD Theory of Mind (ToM) measure, a novel, linguistically simple, tablet-based, and behavioral ToM assessment tool designed for both NT children and children with ASD within an international multi-site project. Given that ToM, the ability to understand that humans can have different perspectives (Premack and Woodruff, 1978), is crucial for social interaction and communication, researchers from different disciplines are interested in the use of a ToM measurement tool that is appropriate for children. Therefore, the first objective of this paper was to synthesize the claims made by reviews of existing ToM measures (Ziatabar Ahmadi et al., 2015; Beaudoin et al., 2020; Quesque and Rossetti, 2020; Osterhaus and Bosacki, 2022; Fu et al., 2023), that led to the highlight of methodological gaps in existing behavioral ToM measures: their high verbal demands and lack of engaging material for younger participants, which both present specific challenges for individuals with special needs, such as autistic children, insufficient number of items, difficulties in the modes of presentation, as well as high reliance on interactions with test administrators. Based on the need for a new behavioral ToM measure, the second objective was to present the developmental process, the structure, and the preliminary psychometric evaluation of the newly created ABCCD ToM measure that was completed in adherence to the manual "Standards for Educational and Psychological Testing". The new ToM measure needed to meet the following criteria: (1) It must be well described with respect to the intended score interpretations. (2) It must be valid. (3) It must be reliable. (4) It must be accessible to children with ASD and NT children; thus, it should not be too long, and it should rely the least possible on comprehension of complex language structures and interaction with the test administrators. The behavioral, linguistically simple, and tablet-based ABCCD ToM measure consists of three blocks with four test items each that were designed to measure the children's abilities to distinguish between their own and someone else's desires (Block 1, Diverse Desires), to measure the children's abilities to attribute a false belief to someone else who is holding a false belief about reality (Block 2, First-order False Beliefs), and to measure the children's abilities to attribute a false belief to someone else who is holding a false belief about a third's person belief (Block 3, Second-order False Beliefs).

4.2 Summary of preliminary findings

The ABCCD ToM measure was preliminarily evaluated with NT and autistic children. Performance on Block 1 was at ceiling for both NT and autistic children, which is comparable with findings from Wellman and Liu (2004) who analyzed five constructs (Diverse Desire, Diverse Belief, Knowledge Access, Contents False Belief, Real-Apparent Emotion). However, performance on Block 2 was descriptively lower in the ABCCD ToM measure in comparison to the corresponding Contents False Belief item. Several factors could account for these discrepancies: First, one major difference is the number of response choices presented. Wellman and Liu (2004) presented two response choices, whereas the ABCCD ToM measure includes three response choices, thereby reducing the likelihood of correct answers by chance, a criterion used in Wellman et al.'s (2001) meta-analysis of developmental ages of ToM. Second, the age group in Wellman and Liu (2004) included children between 5 years and 0 months and 6 years and 6 months, while we examined children between 5 years and 0 months and 5 years and 11 months. This narrower age range could influence the results, as developmental differences within a span of 6 months can be substantial. Third, the demographic characteristics of the participants might contribute to differences in performance, as variations in cultural background or socio-economic status can lead to discrepancies in ToM (Wellman et al., 2001).

The measure presented good internal consistency scores in each block, both within NT and autistic children, measured with the help of Kuder-Richardson 20. The validity assessment was informed by a series of measures. Small correlations were found in both NT and autistic children between the subscale scores of Block 1 and the early subscale scores from the ToMI-2 questionnaire (Hutchins and Prelock, 2016). They may be explained by the fact that the ABCCD ToM behavioral measure only consists of one specific subtype of early ToM development (i.e., diverse desires understanding) assessed with a binary response type. In contrast, the basic subscale of the ToMI-2 questionnaire entails 20 questions scored via a Likert scale encompassing a broader range of early ToM constructs. Therefore, higher correlations were not hypothesized and could not be expected. At the same time, the ToMI-2 presents a questionnaire assessment that may be guided by a "biased" parental judgment. Therefore, the results of the correlations with the subscales of the ToMI-2 for Block 1

should not be overinterpreted. Moderate correlations were found for Block 2 and the basic subscale of the ToMI-2, and Block 3 and the advanced subscale of the ToMI-2 in both groups. The non-significant correlation in the autistic children's group can be explained by the small sample sizes (Block 2: $N = 54$, Block 3: $N = 26$) which may not be high enough to capture small to moderate correlations (Komaroff, 2020). The low and in some cases nonsignificant correlations could be due to small sample sizes and the choice of the measure. In other words, it could be that the correlations would have been higher if we had chosen another ToM measure. Further investigations of validity showed that age, language proficiency and working memory were significant predictors for the performance on the ABCCD ToM measure, for both NT and autistic children. Consequently, the link between the ABCCD ToM measure and other aspects of children's development provides further validity evidence, as noted by Hutchins and Prelock (2016) and Tahiroglu et al. (2014). Additionally, analyses demonstrated that the tool can distinguish between NT and autistic children, offering another source of validity evidence, given that difficulties in ToM can be present in autistic children (e.g., Tager-Flusberg, 2007).

Scale- and item-level analyses of the psychometric properties of the ABCCD ToM measure were completed only within the NT children, due to low sample size within the autistic children's group. EFA showed that the measure assesses within each block a single construct consistent with our design objectives. The application of IRT with the help of Rasch modeling provided a detailed item-level evaluation. It demonstrated that items varied appropriately in difficulty across the latent trait continuum. The results showed that the item difficulties are, within each block, relatively small. This can be explained in two ways: On one hand, the item difficulties in Block 1 were expected to be small since the understanding of diverse desires is considered in NT children to be acquired around the age of 3, which is younger than our target age range for the use of this measure. However, because heterogeneity within autistic samples can be very high, the ABCCD ToM measure was designed to contain this block to situate this easier ToM ability in the target population. Further, the target population includes autistic children who are considered to potentially present difficulties in understanding diverse desires, and this IRT modeling has been applied to NT children who are considered to have fewer difficulties; accordingly, the difficulty range in NT children is generally assumed to be a bit lower. On the other hand, the item difficulties in Blocks 2 and 3 also appear not to be specifically difficult. However, Block 3 measures the ability to attribute a belief to another person with a false belief about a third person's perspective, a relatively challenging condition. Nevertheless, small item difficulties can be explained by the fact that stop criteria were introduced between blocks that may have masked greater difficulty scores. Due to these stop criteria, participants who had scored incorrectly in more than two test items within a block would not have seen the subsequent block. If the following blocks had also been administered to these participants, we would have expected higher difficulty scores. Therefore, these analyses reinforce the measure's robustness and capability to provide precise estimates of diverse desires understanding, first-order false belief attribution, and second-order false belief attribution in NT children between 4 and 10 years of age.

5 Conclusion

Overall, the newly created ABCCD Theory of Mind (ToM) measure offers a promising new tool for the assessment of ToM that is tablet-based, linguistically simple, highly visual, and therefore accessible and engaging for neurotypical (NT) and autistic children between 4 and 10 years. Crucial methodological weaknesses highlighted in reviews of existing ToM measures, such as alignment between the conceptual and methodological definition of ToM, insufficient number of items, and a potential bias introduced through their modes of presentations have been addressed carefully. Furthermore, the preliminary psychometric analysis of the new ToM measure provides insights into the measure's psychometric properties. To allow replication of this validation study with a larger cohort and inclusion of scale- and item-level measures, such as EFA and IRT also in autistic children, all materials to use the ABCCD ToM measure via an application or a manual assessment are made available on OSF. Furthermore, guidelines are presented on how other language versions can be easily created from the existing material.

5.1 Limitations

While the ABCCD ToM measure addresses important gaps in other behavioral measures of ToM, some limitations must be acknowledged. First, the sample size of the autistic children's group was relatively small, and the item- and scale-level analyses with the help of EFA and IRT were therefore only completed within NT children (Morizot et al., 2007), which might limit the generalizability of our findings to those with ASD; therefore, these are preliminary results.

Second, although no significant differences were found between assessment in school vs. at home, the testing environment was not the same for all children which presents a limitation to this study.

Third, the ABCCD ToM measure focuses only on three subdomains of ToM—diverse desires, first-order false beliefs, and second-order false beliefs. Although these subcomponents are critical for assessing fundamental ToM capabilities, other constructs, such as understanding emotions, were not explicitly measured. However, this choice had to be made due to practical time constraints, which pose a challenge in behaviorally assessing the developmental cognitive status of children, particularly those with developmental disorders (American Educational Research Association et al., 2014). Children with autism, in particular, may struggle to sustain attention and remain engaged in tasks (Hours et al., 2022), necessitating a shorter assessment duration. If we had included additional constructs, we might have observed different patterns in the development of ToM skills. For instance, incorporating knowledge access might have highlighted another early developmental stage, in addition to understanding diverse desires. Including understanding emotions might have provided more insight into the social and emotional aspects of ToM. Future research could expand on these constructs to encompass a wider array of ToM skills, offering a more detailed understanding of the development and nuances of ToM across different contexts and populations.

Additionally, many behavioral ToM measures include only one test item per construct and offer two response choices, increasing the likelihood of correct answers by chance (Ziatabar Ahmadi et al., 2015). Furthermore, ToM measures often require interaction between the participant and the test administrator, but most children with ASD struggle with social interaction due to social anxiety (Montaser et al., 2023). Considering these challenges in behaviorally assessing ToM in autistic children, the development of the ABCCD ToM measure was guided by the need to avoid potential misinterpretations of performances. Therefore, the ABCCD ToM measure assesses each of the three constructs with four test items, allowing for more accurate measurement within a timeframe of 20–25 min. This approach prioritized including multiple test items per construct over the number of constructs to ensure valid interpretations of performance. Focusing on cognitive abilities, we included three constructs that reflect important milestones in the development of ToM (Wimmer and Perner, 1983; Wellman and Liu, 2004). If additional constructs, including affective ToM, need to be assessed in children, we agree with other researchers (Merrell, 2007; Beaudoin et al., 2020) that these evaluations should be conducted through parental questionnaires, such as the ToMI-2 (Hutchins et al., 2012) for older children, or the CSUS (Tahiroglu et al., 2014) for younger children. However, parental assessments have limitations, such as difficulties in conducting assessments in low SES families, variability in parents' interpretations of their children's behavior, and potential over- or underestimation of abilities (Tahiroglu et al., 2014). Therefore, it is essential not to rely solely on parental reports but to integrate behavioral tasks and provide thorough instructions for parents on accurately completing these questionnaires. The ToMI-2 (Hutchins et al., 2012), for example, offers guidelines by asking parents to indicate the degree to which they believe the statements are true for their child, accompanied by three examples. A key indicator that parental ToM questionnaires are effective for children with ASD is their ability to discriminate between typical and atypical groups. This was demonstrated for the CSUS, although only on a small sample of 18 NT and 15 autistic children (Tahiroglu et al., 2014), and with the ToMI-2 (Hutchins and Prelock, 2016). Thus, combining direct behavioral assessments with indirect parental assessments of ToM can provide a more comprehensive understanding of children's ToM abilities. Hutchins and Prelock (2016) similarly included the "Theory of Mind Task Battery" in the "Theory of Mind Inventory", to address the need for direct assessment of ToM competencies for various research and clinical purposes.

5.2 Future directions

Future research should consider conducting a validation study with a larger cohort both in NT children as well as in autistic children as for this study no IRT analyses were possible due to a too small sample (Morizot et al., 2007). To allow an advanced assessment of item difficulties within IRT, all three blocks should be administered to all participants even if the stop criterion would have been reached. Furthermore, longitudinal applications may assess the measure's sensitivity over time in individual ToM abilities.

Data availability statement

The datasets presented in this study can be found in OSF: https://osf.io/pg2an/?view_only=ce9836db48d7477db52f79db7187995b.

Ethics statement

The studies involving humans were approved by the Swiss Association of Research Ethics Committees (Swissethics, Switzerland, Project ID-2022-00878) and the Institutional Review Board of the University of Connecticut (USA). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

FB: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. PW: Writing – review & editing, Project administration, Methodology. SS: Writing – review & editing, Formal analysis, Conceptualization. NR: Writing – review & editing, Supervision. MC: Writing – review & editing, Software. MD: Writing – review & editing, Supervision. AS: Writing – review & editing. GC: Writing – review & editing, Methodology. LN: Writing – review & editing, Supervision, Methodology. SD: Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Swiss National Science Foundation (SNSF) with a PRIMA grant awarded to SD (number: PR00P1_193104/1).

Acknowledgments

We warmly thank all participants who participated in this study and all autism centers, teachers, and clinicians for their support in reaching out to participants. We also thank all our collaborators and students, who supported the material creation with valuable feedback and pilot testing, and for supporting the data collection process at various places within Europe and the United States. We thank Jill de Villiers, Inge-Marie Eigsti and the entire MiLA-group for their valuable feedback on initial phases of the material creation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdpys.2024.1445406/full#supplementary-material

References

Alhajeri, O., Anderson, J. A., and Alant, E. (2017). Effectiveness of the use of ipads to enhance communication and learning for students with autism: a systematic review. *IJTIE.* 5:6. doi: 10.20533/ijtie.2047.0533.2017.0132

Alzrayer, N., Banda, D. R., and Koul, R. K. (2014). Use of iPad/iPods with individuals with autism and other developmental disabilities: a meta-analysis of communication interventions. *Rev. J. Autism Dev. Disord.* 1, 179–191. doi: 10.1007/s40489-014-0018-5

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x

Audacity Team (2014). *Audacity(R): Free Audio Editor and Recorder [Computer program]*. Available at: https://www.audacityteam.org/ (accessed April 1, 2022).

Baker, C. (2011). *Foundations of Bilingual Education and Bilingualism*. Bristol: Multilingual Matters.

Baker, F. B. (2002). *The Basics of Item Response Theory, 2nd Edn*. College Park: ERIC Clearinghouse on Assessment and Evaluation.

Baron-Cohen, S. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.

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

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* 67, 1–47. doi: 10.18637/jss.v067.i01

Bean, G. J., and Bowen, N. K. (2021). Item response theory and confirmatory factor analysis: complementary approaches for scale development. *J. Evid.-Based Soc. Work* 18, 597–618. doi: 10.1080/26408066.2021.1906813

Beaudoin, C., Leblanc, É., Gagner, C., and Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. *Front. Psychol.* 10:2905. doi: 10.3389/fpsyg.2019.02905

Bottema-Beutel, K., Kapp, S. K., Lester, J. N., Sasson, N. J., and Hand, B. N. (2021). Avoiding ableist language: suggestions for autism researchers. *Autism Adulth.* 3, 18–29. doi: 10.1089/aut.2020.0014

Broekhof, E., Ketelaar, L., Stockmann, L., Van Zijp, A., Bos, M. G. N., and Rieffe, C. (2015). The understanding of intentions, desires and beliefs in young children with autism spectrum disorder. *J Autism Dev Disord.* 45, 2035–2045. doi: 10.1007/s10803-015-2363-3

Buijsman, R., Begeer, S., and Scheeren, A. M. (2023). 'Autistic person' or 'person with autism'? Person-first language preference in Dutch adults with autism and parents. *Autism* 27, 788–795. doi: 10.1177/13623613221117914

Burnel, M., Perrone-Bertolotti, M., Reboul, A., Baciu, M., and Durrleman, S. (2018). Reducing the language content in ToM tests: a developmental scale. *Dev. Psychol.* 54, 293–307. doi: 10.1037/dev0000429

Carlson, S. M., Moses, L. J., and Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: an investigation of inhibitory control and planning ability. *J. Exp. Child Psychol.* 87, 299–319. doi: 10.1016/j.jecp.2004.01.002

Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the *R* environment. *J. Stat. Soft.* 48, 1–29. doi: 10.18637/jss.v048.i06

De Cat, C., Kašćelan, D., Prevost, P., Serratrice, L., Tuller, L., and Unsworth, S. (2022). *Quantifying Bilingual EXperience (Q-BEx): Questionnaire Manual and Documentation*.

De Villiers, J., and Pyers, J. (2002). Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cogn. Dev.* 17, 1037–1060. doi: 10.1016/S0885-2014(02)00073-4

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *BBS* 1, 568–570. doi: 10.1017/S0140525X00076664

Dore, R. A., Amendum, S. J., Golinkoff, R. M., and Hirsh-Pasek, K. (2018). Theory of mind: a hidden factor in reading comprehension? *Educ. Psychol. Rev.* 30, 1067–1089. doi: 10.1007/s10648-018-9443-9

Dunn, L. M., and Dunn, D. M. (2007). *PPVT-4: Peabody Picture Vocabulary Test*. San Antonio: Pearson Assessments.

Dunn, L. M., Dunn, L. M., Stella, G., Pizzoli, C., and Tressoldi, P. (2016). *PPVT-Revised: Peabody Picture Vocabulary Test [Italian adaptation]*. Torino: Omega Edition.

Dunn, L. M., Dunn, L. M., and Thériault-Whalen, C. M. (1993). *Echelle de vocabulaire en Images Peabody: EVIP*. Toronto: PSYCAN.

Durrleman, S., Marinis, T., and Franck, J. (2016). Syntactic complexity in the comprehension of wh-questions and relative clauses in typical language development and autism. *Appl. Psycholinguist.* 37, 1501–1527. doi: 10.1017/S0142716416000059

Forgeot d'Arc, B. F., and Ramus, F. (2011). Belief attribution despite verbal interference. *Q J Exp. Psychol. B* 64, 975–990. doi: 10.1080/17470218.2010.524413

Fu, I.-N., Chen, K.-L., Liu, M.-R., Jiang, D.-R., Hsieh, C.-L., and Lee, S.-C. (2023). A systematic review of measures of theory of mind for children. *Dev. Rev.* 67, 101061. doi: 10.1016/j.dr.2022.101061

Georgopoulos, M. A., Brewer, N., Lucas, C. A., and Young, R. L. (2022). Speed and accuracy of emotion recognition in autistic adults: the role of stimulus type, response format, and emotion. *Autism Res.* 15, 1686–1697. doi: 10.1002/aur.2713

Haas, J. K. (2014). *A History of the Unity Game Engine*.

Hours, C., Recasens, C., and Baleyte, J.-M. (2022). ASD and ADHD comorbidity: what are we talking about? *Front. Psychiatry* 13:837424. doi: 10.3389/fpsyt.2022.837424

Hutchins, T. L., and Prelock, P. A. (2016). *Technical Manual for the Theory of Mind Inventory-2*. Unpublished Copyrighted Manuscript. Available at: theoryofmindinventory.com (accessed December 1, 2021).

Hutchins, T. L., Prelock, P. A., and Bonazinga, L. (2012). Psychometric evaluation of the theory of mind inventory (ToMI): a study of typically developing children and children with autism spectrum disorder. *J. Autism Dev. Disord.* 42, 327–341. doi: 10.1007/s10803-011-1244-7

Hutchins, T. L., Prelock, P. A., and Chace, W. (2008). Test-retest reliability of a theory of mind task battery for children with autism spectrum disorders. *Focus Autism Other Dev. Disabl.* 23, 195–206. doi: 10.1177/1088357608322998

Jacobs, J., and Paris, S. (1987). Children's metacognition about reading: issues in definition, measurement, and instruction. *Educ. Psychol.* 22, 255–278. doi: 10.1080/00461520.1987.9653052

Joseph, R. M., and Tager-Flusberg, H. (2004). The relationship of theory of mind and executive functions to symptom type and severity in children with autism. *Dev. Psychopathol.* 16, 137–155. doi: 10.1017/S095457940404444X

Kenny, D. A., Kaniskan, B., and McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociol. Methods Res.* 44, 486–507. doi: 10.1177/0049124114543236

Komaroff, E. (2020). Relationships between p-values and pearson correlation coefficients, type 1 errors and effect size errors, under a true null hypothesis. *J. Stat. Theory Pract.* 49, 1–13. doi: 10.1007/s42519-020-00115-6

Kuder, G. F., and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika* 2, 151–160. doi: 10.1007/BF02288391

Lenhard, A., Lenhard, W., Segerer, R., and Suggate, S. (2015). *Peabody Picture Vocabulary Test-4.* Deutsche Fassung: Pearson Assessment.

Lord, C., Rutter, M., DiLavore, P. C., and Risi, S. (2003). *Autism Diagnostic Observation Schedule: ADOS*. Los Angeles, CA: Western Psychological Services.

Lord, C., Rutter, M., and Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* 24, 659–685. doi: 10.1007/BF02172145

Marinis, T., Andreou, M., Bagioka, D. V., Baumeister, F., Bongartz, C., Czypionka, A., et al. (2023). Development and validation of a task battery for verbal and non-verbal first- and second-order theory of mind. *Front. Lang. Sci.* 1, 1–17. doi: 10.3389/flang.2022.1052095

Marocchini, E. (2023). Impairment or difference? The case of theory of mind abilities and pragmatic competence in the autism spectrum. *Appl. Psycholinguist.* 44, 365–383. doi: 10.1017/S0142716423000024

Merrell, K. W. (2007). *Behavioral, Social, and Emotional Assessment of Children and Adolescents, 3rd Edn*. New York: Routledge.

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychol. Bull.* 135, 749–773. doi: 10.1037/a0016854

Milligan, K., Astington, J. W., and Dack, L. A. (2007). Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child Dev.* 78, 622–646. doi: 10.1111/j.1467-8624.2007.01018.x

Montaser, J., Umeano, L., Pujari, H. P., Nasiri, S. M. Z., Parisapogu, A., Shah, A., et al. (2023). Correlations between the development of social anxiety and individuals with autism spectrum disorder: a systematic review. *Cureus* 15:e44841. doi: 10.7759/cureus.44841

Morizot, J., Ainsworth, A. T., and Reise, S. P. (2007). "Toward modern psychometrics: Application of item response theory models in personality research," in *Handbook of Research Methods in Personality Psychology*, eds. R. W. Robins, R. C. Fraley, and R. F. Krueger (New York City, NY), 407–423.

Naigles, L. R. (2021). It takes all kinds (of information) to learn a language: investigating the language comprehension of typical children and children with autism. *Curr. Dir. Psychol. Sci.* 30, 11–18. doi: 10.1177/0963721420969404

Ntumi, S., Agbenyo, S., and Bulala, T. (2023). Estimating the psychometric properties (item difficulty, discrimination and reliability indices) of test items using Kuder-Richardson approach (KR-20). *SIJEd* 11, 18–28. doi: 10.34293/education.v11i3.6081

Osterhaus, C., and Bosacki, S. L. (2022). Looking for the lighthouse: a systematic review of advanced theory-of-mind tests beyond preschool. *Dev. Rev.* 64:101021. doi: 10.1016/j.dr.2022.101021

Perner, J., and Wimmer, H. (1985). "John thinks that Mary thinks that$ldots$" attribution of second-order beliefs by 5- to 10-year-old children. *J. Exp. Child Psychol.* 39, 437–471. doi: 10.1016/0022-0965(85)90051-7

Peterson, C. C., and Wellman, H. M. (2019). Longitudinal theory of mind (ToM) development from preschool to adolescence with and without ToM delay. *Child Dev.* 90, 1917–1934. doi: 10.1111/cdev.13064

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *BBS* 1, 515–526. doi: 10.1017/S0140525X00076512

Quesque, F., and Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspect. Psychol. Sci.* 15, 384–396. doi: 10.1177/1745691619896607

R Core Team (2020). *R Core Team (2020).* Vienna: A Language and Environment for Statistical Computing.

Rapin, I., and Dunn, M. (2003). Update on the language disorders of individuals on the autistic spectrum. *Brain Dev-JPN.* 25, 166–172. doi: 10.1016/S0387-7604(02)00191-2

Raven, J. C., Rust, J., Chan, F., and Zhou, X. (2018). *Raven's Progressive Matrices 2, Clinical Edition*. London: Pearson.

Reise, S. P., Ainsworth, A. T., and Haviland, M. G. (2005). Item response theory: fundamentals, applications, and promise in psychological research. *Curr. Dir. Psychol. Sci.* 14, 95–101. doi: 10.1111/j.0963-7214.2005.00342.x

Roberts, J. A., Rice, M. L., and Tager-Flusberg, H. (2004). Tense marking in children with autism. *Appl. Psycholinguist.* 25, 429–448. doi: 10.1017/S0142716404001201

Röschel, A., Wagner, C., and Dür, M. (2021). Examination of validity, reliability, and interpretability of a self-reported questionnaire on Occupational Balance in Informal Caregivers (OBI-Care) – A Rasch analysis. *PLoS ONE* 16:e0261815. doi: 10.1371/journal.pone.0261815

Rosseel, Y. (2012). lavaan: an *R* package for structural equation modeling. *J. Stat. Soft.* 48, 1–36. doi: 10.18637/jss.v048.i02

Rutter, M., Bailey, A., Berument, S. K., and Lord, C. (2003). "Social communication questionnaire: manual," in *Western Psychological and Counseling Services* (Los Angeles, CA).

Schaeffer, J., Abd El-Raziq, M., Castroviejo, E., Durrleman, S., Ferré, S., Grama, I., et al. (2023). Language in autism: domains, profiles and co-occurring conditions. *J. Neural Transm.* 130, 433–457. doi: 10.1007/s00702-023-02592-y

Shi, D., DiStefano, C., Maydeu-Olivares, A., and Lee, T. (2022). Evaluating SEM model fit with small degrees of freedom. *Multivariate Behav. Res.* 57, 179–207. doi: 10.1080/00273171.2020.1868965

Silleresi, S. (2023). *Developmental Profiles in Autism Spectrum Disorder*. Amsterdam: John Benjamins Publishing Company.

Slaughter, V., Imuta, K., Peterson, C. C., and Henry, J. D. (2015). Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Dev.* 86, 1159–1174. doi: 10.1111/cdev.12372

Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., et al. (2012). Categorical and dimensional definitions and evaluations of symptoms of adhd: history of the SNAP and the SWAN rating scales. *Int. J. Educ. Psychol.* 10, 51–70.

Tager-Flusberg, H. (2007). Evaluating the Theory-of-Mind Hypothesis of Autism. *Curr. Dir. Psychol. Sci.* 16, 311–315. doi: 10.1111/j.1467-8721.2007.00527.x

Tager-Flusberg, H., and Sullivan, K. (1994). Predicting and explaining behavior: a comparison of autistic, mentally retarded and normal children. *J. Child Psychol. Psychiat. Allied Discipl.* 35, 1059–1075. doi: 10.1111/j.1469-7610.1994.tb01809.x

Tahiroglu, D., Moses, L. J., Carlson, S. M., Mahy, C. E. V., Olofson, E. L., and Sabbagh, M. A. (2014). The Children's Social Understanding Scale: construction and validation of a parent-report measure for assessing individual differences in children's theories of mind. *Dev. Psychol.* 50, 2485–2497. doi: 10.1037/a0037914

Vivanti, G. (2020). Ask the editor: what is the most appropriate way to talk about individuals with a diagnosis of autism? *J. Autism. Dev. Disord.* 50, 691–693. doi: 10.1007/s10803-019-04280-x

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

Wellman, H. M., and Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Dev.* 75, 523–541. doi: 10.1111/j.1467-8624.2004.00691.x

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

World Health Organization (2019). *International Statistical Classification of Diseases and Related Health Problems, 11th Edn*. Available at: hQps://icd.who.int/browse/2024-01/mms/en

Ziatabar Ahmadi, S. Z., Jalaie, S., and Ashayeri, H. (2015). Validity and reliability of published comprehensive theory of mind tests for normal preschool children: a systematic review. *Iran. J. Psychiatry* 10, 214–224.