frontiers | Frontiers in Language Sciences

TYPE Methods
PUBLISHED 09 January 2023
DOI 10.3389/flang.2022.1052095

OPEN ACCESS

EDITED BY
Moreno I. Coco,
Sapienza University of Rome, Italy

REVIEWED BY
Giuditta Smith,
University of East Anglia,
United Kingdom
Laura M. Morett,
University of Alabama, United States

*CORRESPONDENCE
Theodoros Marinis
✉ t.marinis@uni-konstanz.de

SPECIALTY SECTION
This article was submitted to Psycholinguistics, a section of the journal Frontiers in Language Sciences

RECEIVED 23 September 2022
ACCEPTED 14 December 2022
PUBLISHED 09 January 2023

CITATION
Marinis T, Andreou M, Bagioka DV, Baumeister F, Bongartz C, Czypionka A, Golegos A, Peristeri E, Skrimpa V, Durrleman S and Terzi A (2023) Development and validation of a task battery for verbal and non-verbal first- and second-order theory of mind.
Front. Lang. Sci. 1:1052095.
doi: 10.3389/flang.2022.1052095

Development and validation of a task battery for verbal and non-verbal first- and second-order theory of mind

Theodoros Marinis[1]*, Maria Andreou[2], Dafni Vaia Bagioka[3], Franziska Baumeister[4], Christiane Bongartz[5], Anna Czypionka[1], Angelika Golegos[1], Eleni Peristeri[6], Vasileia Skrimpa[5], Stephanie Durrleman[4] and Arhonto Terzi[3]

[1]Department of Linguistics, University of Konstanz, Konstanz, Germany, [2]Department of Speech & Language Therapy, University of Peloponnese, Kalamata, Greece, [3]Department of Speech & Language Therapy, University of Patras, Patras, Greece, [4]Autism, Bilingualism, Cognitive and Communicative Development (ABCCD) Lab, Faculty of Science and Medicine, University of Fribourg, Fribourg, Switzerland, [5]English Department, University of Cologne, Cologne, Germany, [6]Department of Theoretical & Applied Linguistics, School of English, Aristotle University of Thessaloniki, Thessaloniki, Greece

This paper presents a new toolkit for assessing Theory of Mind (ToM) *via* performance in first and second-order false belief (FB) tasks. The toolkit includes verbal and non-verbal versions of first and second-order FB tasks; the verbal version is currently available in Greek and German. Scenarios in the toolkit are balanced for factors that may influence performance, like the reason for the FB (deception, change-of-location, unexpected content). To validate our toolkit, we tested the performance of neurotypical adults in the non-verbal and verbal versions in two studies: Study 1 with 50 native speakers of German and Study 2 with 50 native speakers of Greek. The data from both studies yield similar results. Participants performed well in all conditions, showing slightly more difficulties in the second- than first-order FB conditions, and in the non-verbal than the verbal version of the task. This suggests that the task is at the high end of the sensitive range for neurotypical adults, and is expected to be well inside the sensitive range for children and populations that have difficulties in ToM. Factors like deception and type of outcome in the video-scenarios did not influence the behavior of neurotypical adults, suggesting that the task does not have any confounds related to these factors. The order of presentation of the verbal and non-verbal version has an influence on performance; participants beginning with the verbal version performed slightly better than participants beginning with the non-verbal version. This suggests that neurotypical adults used language to mediate ToM performance and learn from a language-mediated task when performing a non-verbal ToM task. To conclude, our results show that the scenarios in the toolkit are of comparable difficulty and can be combined freely to match demands in future research with neurotypical children and autistic individuals, as well as other populations that have been shown to have difficulties in ToM.

Differences between baseline and critical conditions can be assumed to reflect ToM abilities, rather than language and task-based confounding factors.

KEYWORDS

autism, Theory of Mind, false belief, verbal, non-verbal, adults, first-order, second-order

1. Introduction

Theory of Mind (ToM) is the mental ability which allows people to understand that other people have beliefs, desires and intentions that differ from their own, and to predict ensuing behaviors (Premack and Woodruff, 1978; Flobbe et al., 2008). ToM is characterized as a multi-component ability that includes social understanding, emotional mental state recognition, perspective taking, and sarcasm. Thus, it strongly influences the quality of people's communication, as it provides individuals with an implicit social "know-how" that allows them to negotiate everyday social interactions. The ability to negotiate the mental domain may, in turn, enable empathic exchanges and joint goal-directed activity that promotes communication and social competence in general. Furthermore, ToM is a multifaceted process that encompasses the collaboration of a variety of cognitive functions, such as attention, working memory, and language comprehension, and relies on the operational development of the brain (Korkmaz, 2011). ToM skills start from a child's reasoning from their own point of view to taking into consideration another person's beliefs (first-order), and later to taking into consideration another person's beliefs about other person's beliefs (second-order) (Wellman and Liu, 2004).

ToM deficits have long been regarded as one of the most disabling features in Austism Spectrum Disorder (ASD), a neurodevelopmental disorder that includes difficulties in communication and social interaction. Autistic individuals demonstrate deviations between the perception of their own knowledge and that of others (Baron-Cohen et al., 1985), a fact that leads to poor performance in ToM tasks (Zelazo et al., 2002; Andreou and Skrimpa, 2020). Although correlations have been found between their performance in ToM tasks and their social competence, individuals within the spectrum have been shown to have difficulties in implementing ToM within social contexts, despite the fact that they can regenerate thoughts, beliefs and intentions in ToM tasks (Begeer et al., 2011; Chevallier et al., 2014). Multiple lines of evidence point to difficulties in representing mental states in autism. Autistic children have been shown to have difficulties in ToM tasks that tap into other-referential cognition (Bodner et al., 2015) which may underpin social functioning and communication deficits. Studies have also shown that autistic adults can have challenges in responding to inference questions about others' mental states when processing pragmatically complex social scenarios (Heavey et al., 2000; Lönnqvist et al., 2017) or they tend to engage less in inferring the broader meaning of social events when telling narratives (Barnes and Baron-Cohen, 2012).

Despite extensive research on topics of social cognition in autistic individuals, investigations targeting the vulnerability of this population to deception detection have largely relied on the assessment of first-order ToM tasks. Turning to child studies, unexpected change of location and unexpected change of contents have so far been the most extensively studied constrained situations designed specifically to elicit false belief (FB) attribution in younger autistic children (Baron-Cohen et al., 1985; Andreou et al., 2020; Baldimtsi et al., 2021; Peristeri et al., 2021; Durrleman et al., 2022a).

The Sally-Ann task (Wimmer and Perner, 1983) has been a standardized way of assessing first-order ToM of autistic individuals in an explicit manner.[1] In this task, Sally leaves her marble in the basket not having witnessed the fact that Ann has put it in the box, while Sally was out of the room. The right answer to the question of where Sally will look for her marble is the marble's original location (the basket). Autistic children face difficulties with this task and respond that Sally will look for the marble in the box and not in the basket. This difficulty demonstrates a genuine inability to understand that other people have different beliefs from their own (Baron-Cohen, 1995). Mentalizing abilities and deviations in both neurotypical and autistic individuals have typically been investigated using similar versions of the image- or puppet-based Sally-Anne task, whereby holding a FB about an event in the world has a mnemonic advantage over holding a FB about someone else's belief (i.e., second-order ToM), the latter representing higher-order recursive mentalistic reasoning (Lecce et al., 2014; Arslan et al., 2017). Moreover, first-order ToM tests have often employed static, non-interactive images and puppets, thus,

1 In the present study, we restrict our scope to explicit ToM tasks, namely, tasks that require an explicit response, in contrast to implicit ToM tasks, e.g., tasks using eye-tracking. This is because explicit tasks can be used in clinical and educational settings because they are easier to administer and analyze. Implicit tasks are mainly used for research purposes because they require lengthy data analyses to evaluate success in ToM.

neglecting the complex and dynamic character of the real world, which involves extracting information from incoming sensory input, developing predictions regarding the timing of upcoming events, and flexibly updating predictions based on the behavior of real-life agents.

The new online toolkit for assessing ToM fills this gap by addressing both first- and second-order FB attribution abilities, using interactive, real-life scenarios depicted through video stimuli. In this manner, the toolkit is suitable to investigate basic-level/first-order as well as advanced/second-order ToM skills in neurotypical and autistic individuals, as well as other populations with vulnerabilities in social interactions.

1.1. The role of language in ToM

The interface between ToM and language perception and production appears to be bidirectional. On one hand, the emergence of ToM enables language development at an early age (Nelson, 2005); on the other hand, language skills (in terms of lexicon, morphosyntax, and pragmatics) promote ToM abilities (Harris et al., 2005; Milligan et al., 2007; Andreou et al., 2020; Durrleman et al., 2022b), as indicated by children's performance in FB ToM tasks. More specifically, language ability, including vocabulary and morpho-syntactic skills, has been shown to be a powerful predictor of ToM performance in neurotypical children (de Villiers and de Villiers, 2014). In particular, syntax seems to be a significant component in the attribution of beliefs and the development of ToM; especially, embedded syntax [e.g., *Sally thinks (that the ball is in the basket)*], which is also often used in FB tests, mostly due to the inclusion of verbs that reflect beliefs or mental states of the agent in the main clause (de Villiers and de Villiers, 2009). The aforementioned studies show that language competence expressed in either syntactic embedding or knowledge of the factivity semantics of mental terms correlates with neurotypical children's ToM reasoning.

Autistic children have been reported to present delays in language acquisition as well as difficulties within the domains of pragmatics and prosody (Lord et al., 2004; Roberts et al., 2004; Chevallier et al., 2010; Diehl et al., 2015, among others). At the same time, they have been found to perform poorly at ToM tasks since Baron-Cohen et al. (1985). Research on language as a predictor of social cognition difficulties in autistic children argues that individuals in the spectrum can often benefit from specific components of language, such as morphosyntax, in order to adequately perform in ToM tasks, i.e., they build upon morphosyntactic experience as an adaptive means of scaffolding ToM, and specifically, FB reasoning (Happé, 1995; Tager-Flusberg, 2000; Durrleman et al., 2022b). This mechanism comprises the computation of FB attributions which compensates for the deficient mental representations of the perspectives of others. However, the overwhelming majority of studies that have explored ToM skills in autistic children have been conducted in English; there is thus a shortage in the availability of data from autistic children speaking other languages than English, as well as in the availability of tools measuring ToM in languages other than English.

1.2. Low-verbal ToM tasks in autism

Besides the lack of studies of ToM in autism in languages other than English, what has been confounded in previous research is the ability to tease apart autistic individuals' ToM skills and their language competence. Although language delays in autism have provided researchers with the opportunity to test for possible correlations between ToM and language competence, the fact is that it is very challenging to find sets of measures that purely tap into ToM and yield sufficient variance. This is because the most frequently used ToM measures either require an advanced vocabulary (e.g., Norbury and Bishop, 2003; Peristeri et al., 2017) or involve comprehension of verbal scenarios of around 100 words (e.g., Sally Anne test; Wimmer and Perner, 1983). This burdens language comprehension, verbal short-term and working memory, and further asks participants to formulate responses to fairly complex questions (e.g., *When Sally returns, where will she look for her marble*?). Difficulties to perform well in these tasks could be related to the participants' verbal rather than mentalizing ability or to the large heterogeneity that characterizes individuals with autism (e.g., Roberts et al., 2004; Andreou and Skrimpa, 2022).

Few studies to date have attempted to tease apart these factors in children by using low-verbal ToM experimental paradigms. Peristeri et al. (2021) tested first-order ToM skills in 7- to 15-year old monolingual and bilingual autistic children through an online low-verbal task comprising 19 cartoon-based video scenarios (adapted from Forgeot d'Arc and Ramus, 2011). Each scenario involved a change that was witnessed or not by a main character/protagonist. Each participant had to decide with a Yes/No response about the appropriateness of the scenario's ending, which in half of the trials aligned with the belief of the main character (which was different from the participant's). Control items involved deciding on the appropriateness of the scenario's ending due to change in the physical world (i.e., causal relations). Despite the low-verbal format of the task, Peristeri et al. (2021) found that monolingual autistic children exhibited lower accuracy in attributing FB to the character of the scenario as compared to their age- and SES-matched bilingual peers, although they performed well on control items. Durrleman et al. (2016) also tested first-order FB skills *via* a low-verbal task. The study involved a picture-sequencing-task (adapted from Baron-Cohen et al., 1985) administered to children with ASD from 6 to 14 years old. In order to arrive at the correct sequence for the critical test-items, children had to understand intentions and reactions of a character in relation to their FB, while successful performance on control

items again depended on children's grasp of causal relations. As in the study by Peristeri et al. (2021), children found causal sequences easier than the sequences involving belief reasoning. Interestingly, the ability to succeed on the latter was related to the children's linguistic skills, and in particular on their mastery of embedded propositions, along the lines of that found by Hollebrandse et al.'s (2014) for younger neurotypicals with a more complex ToM task. Indeed, in Hollebrandse et al. (2014), 6-to-9-year-old neurotypical children were administered not only first- but also second-order verbal ToM tasks with stories and corresponding low-verbal ToM tasks whose stimuli included short video clips. According to their findings, while children managed to perform near-ceiling level in first-order ToM tasks irrespective of the type (verbal vs. non-verbal) of the stimuli, a clear performance advantage was observed for the verbal (vs. non-verbal) second-order ToM tasks. According to Hollebrandse et al. (2014), these findings provide evidence that language may act as a facilitator of advanced ToM reasoning by providing a means for hierarchically embedding propositions within other propositions which is, in turn, isomorphic to the children's ability to embed other agents' mental representations into their own.

The aforementioned low-verbal ToM paradigms have been useful in paving the way for designing less word-heavy measures of second-order FB attribution skills. However, limitations continue to exist, since studies by Peristeri et al. (2021) and Durrleman et al. (2016) included first-order ToM scenarios only, while Hollebrandse et al.'s (2014) low-verbal second-order ToM tasks involved questions (e.g., "What does Sam think they are selling at the bake sale? What will Maria say to the mailman?") that may have burdened children's mentalizing performance. The current study addresses these limitations by proposing a comprehensive toolkit that tests first- and second-order ToM in the absence of language demands, while at the same time providing the possibility to test ToM skills in a parallel way in the verbal modality.

2. The aims of the new toolkit

To address the gaps and shortcomings of the existing tasks measuring ToM, our first aim was to develop a comprehensive toolkit that tests first- and second-order ToM in a task that uses minimal language for instruction and no language while performing it. Therefore, the task can be appropriate not only for individuals with high but also for those with low verbal abilities. Given that no non-verbal tool is available for the assessment of second-order ToM, there was a clear need for developing such a tool. To be able to compare first with second-order ToM, the toolkit should have a first and a second-order version in a similar format so that the two are directly comparable.

As mentioned already, assessing FB ToM involves different scenario types (e.g., change of location/unexpected content and deception/non-deception), but these types are often not equally represented in the existing tasks. To address this shortcoming, the second aim of the toolkit was to include an equal number of scenarios for each one of the relevant FB types, i.e., change of location/unexpected content and deception/non-deception. Importantly, we have included a good number of scenarios for each type (see Methodology for a description of the types). This enables us to evaluate how people react to different scenario types and the extent to which the reason for FB reasoning affects ToM performance across diverse experimental populations. Including a large number of scenarios also enables us to take out scenarios that may be problematic, and will allow future researchers who use the toolkit to select the scenarios that they consider more appropriate for their studies.

The third aim was to create a toolkit that would be appropriate for both adults and children, so that it can be of use to a variety of researchers and/or clinicians, who may need background assessments of ToM or deeper insights into the language and cognitive skills of different ranges of participants. The toolkit may be valuable for researchers testing not only autistic individuals, but also populations whose language and ToM skills may be impaired, such as brain-damaged individuals (Balaban et al., 2016) and individuals with mental disorders, such as schizophrenia (e.g., Frith and Corcoran, 1996).

Last but not least, we aimed at designing the toolkit for speakers of more than one language other than English in order to make it accessible to a larger number of people as well as to multilingual speakers. To this end, we created the toolkit for German and Greek speakers. Whilst developing the material, we took into consideration a number of cultural similarities and differences the two languages and cultures share, and this makes the toolkit appropriate for two European societies that differ in various ways. Therefore, it is expected that the tool can be easily adapted to other languages and cultures that are similar to either one of the two.

3. The toolkit

The toolkit consists of a non-verbal and a verbal ToM task investigating first- and second-order FB ToM using short video clips and is inspired by Forgeot d'Arc and Ramus (2011). Ninteen scenarios were created, three of which were used for practice and 16 for the main part.

3.1. Types of scenarios

Half of the scenarios in the main task involved an unexpected content and were modeled according to the Smarties task (Wimmer and Perner, 1983). The other half involved a change of location and were created according to the Sally-Ann task (Baron-Cohen et al., 1985).

TABLE 1 Types of scenarios in the main task.

Scenarios with deception		Scenarios without deception	
Scenarios with unexpected content	Scenarios with change of location	Scenarios with unexpected content	Scenarios with change of location
Scenario 1: Egg	Scenario 5: Plants	Scenario 9: Bottle	Scenario 13: Scarf
Scenario 2: Smarties	Scenario 6: Hide and seek	Scenario 10: Dry pen	Scenario 14: Sandwich
Scenario 3: Pencil case	Scenario 7: Papers to glue	Scenario 11: Pizza box	Scenario 15: Lemonade
Scenario 4: Schoolbag	Scenario 8: Toys	Scenario 12: Socks and pencils	Scenario 16: Toy cars

In the scenarios of the type "Unexpected content" there is a change in the scene so that a familiar container (e.g., a pencil case in Scenario 3 "Pencil case") contains an unexpected content (e.g., chocolate eggs instead of pencils). In scenarios of the type "Change of location," an object (e.g., a plant in Scenario 5 "Plants") is moved from its original location (e.g., right on the table) to another location (e.g., left on the table) so that it is no longer in its original location. In half of the scenarios, the change of location or the source of the unexpected content was motivated by deception, whereas in the other half there was no deception. Scenarios with deception have been shown to be easier than scenarios in which the change is not motivated by an intentional deception because deception enables participants to adopt the perspective of the deceived person and to recognize that they hold a FB (Wellman et al., 2001). The manipulation of deception/no deception and unexpected content/change of location gave rise to four scenarios by type, as illustrated in Table 1. Based on these scenarios, short (1-to-2 min) video clips were created with real people in locations within a house.

3.2. Structure of the video clips

The video clips had the same structure. In all of them there is an "outer" scene with a person ("observer") sitting on a chair in front of a TV screen, watching an "inner video." The TV screen shows the "inner" video sequence. Each "inner video" includes two actors ("actor1" and "actor2") who act-out the scenario. Each "inner video" is composed of four phases: beginning, change, suspense, and an end, as shown in Table 2.

Phase 1 (the beginning phase) sets up the scene and introduces actor1. Actor1 starts to perform an action but is interrupted and leaves the scene. For example, in Scenario 15 ("Lemonade," shown in Table 2), Nick (actor1) enters the kitchen takes a bottle of lemonade but he realizes that he doesn't have a bottle opener. He puts the lemonade on a cabinet and leaves the kitchen to look for one. In Phase 2 (the change phase), Anna (actor2) enters the room and sees the lemonade on the cabinet. She puts it into the fridge and leaves the room. In Phase 3 (the suspense phase), actor1 re-enters the kitchen ready to perform the action he started in Phase 1. In Scenario 15, Nick comes back with a bottle opener in order to open the lemonade. In Phase 4 (the end phase), the scene freezes and two alternative endings are presented in the form of two thought bubbles that pop up over the head of the actor1 (for first-order) or the observer (for second-order). The two thought bubbles illustrate two possible actions actor1 may perform next. One will be the appropriate, depending on whether or not actor1 (for first-order) and the observer (for second-order) have witnessed the change. In Scenario 15, one thought bubble shows Nick going toward the cabinet, while the second shows Nick opening the fridge. In the end phase, participants had to select the appropriate ending for the story by clicking with the mouse on one of the two thought bubbles.

3.3. Conditions

To address first- and second-order ToM in each of the 16 scenarios, six different versions of each video clip (six conditions) were created: two first-order video clips (Condition 1 and 2) and four second-order video clips (Condition 3–6). Depending on each condition, different characters of the scenarios (observer, actor1) were present or absent during the second phase (Change). This gives rise to two control conditions (Condition 1 and 3), two experimental conditions (Condition 2 and 5), and two fillers (Condition 4 and 6; see Table 3).

In Conditions 1 and 3, the change is witnessed by the observer and by actor1. These are the control conditions that test True Belief (TB). Condition 1 controls for the participants' ability to attribute a TB to actor1. This means that the

TABLE 2 Structure of each video clip with four phases.

Phase 1: Beginning	Phase 2: Change	Phase 3: Suspense	Phase 4: End
			

TABLE 3 Conditions.

Condition	First-order		Second-order			
	1	2	3	4	5	6
Type	Control (TB)	Experimental (FB)	Control (TB)	Filler	Experimental (FB)	Filler
Observer	Seen	Seen	Seen	Seen	Unseen	Unseen
Actor1	Seen	Unseen	Seen	Unseen	Seen	Unseen

participants have the same belief as actor1 about what is expected to happen next in the scenario. Condition 3 controls for the participants' ability to attribute a TB to the observer. This means that the participants have the same belief as the observer about what is expected to happen next in the scenario. In Condition 2, the change is witnessed by the observer, but it is unseen by actor1; this experimental condition tests first-order FB, and more specifically, the participants' ability to attribute a FB to actor1, i.e., that the participants have a different belief than actor1 about what actor1 is expected to do next. In Condition 5, the change is unseen by the observer and seen by actor1; it is an experimental condition that tests second-order FB, and more specifically the participants' ability to attribute a FB to the observer who in turn attributes a TB to actor1. This means that the participants have a different belief than the observer about what they think actor1 is expected to do next. Conditions 4 and 6 are fillers that were created to balance the design. In Condition 4, the change is seen by the observer and unseen by actor1. In Condition 6, the change is unseen by both the observer and by actor1.

3.4. Non-verbal and verbal task

A non-verbal and a verbal version were created for each video clip that gave rise to a non-verbal and a verbal task. In the non-verbal task, there was no language throughout the task. The video clips were presented silently; the end phase was also silent and the participants had to select the correct ending of the story by clicking on one of the two thought bubbles. Written instructions were presented only at the beginning of the task, guiding participants to watch video clips and select the correct end of the story (see the Procedure below). The same instructions were also presented in the verbal task. However, in the verbal task the video clips were accompanied by narration in German or Greek. In both languages, the narration included sentences of low linguistic complexity so that it would be easy to comprehend by individuals with language deficits, such as autistic children (Burnel et al., 2018). As such, narration included high-frequency vocabulary (avoiding mental state words) (Tager-Flusberg and Sullivan, 1994), present tense verbs (Roberts et al., 2004), main clauses (Durrleman and Delage, 2020), and proper names (Durrleman et al., 2016; Terzi et al., 2016). The test question at the end of the video clips in the verbal version included a temporal marker ("now"/"first"), as in Durrleman et al. (2016), to ensure that participants are asked what they think actor1 will do next, not what actor1 should do next (Siegal and Beattie, 1991).

3.5. Experimental lists

The 16 experimental scenarios in the six conditions gave rise to 96 non-verbal and 96 verbal video clips (total 192 video clips). These were divided into eight non-verbal and eight verbal lists. Each list included each scenario only once (16 scenarios), two video clips for Condition 1 and 2, and three video clips for Conditions 3, 4, 5, and 6.

3.6. Material creation

The video clips were created in private households by six people naive to acting and were filmed with the camera of a mobile phone. Care was taken to create a minimal background in each video clip by including only objects that were necessary for the story and by avoiding using objects that could distract the attention of the participants. The verbal narrations for both

German and Greek were recorded by male native speakers of German and Greek, respectively. The German narrations were recorded with a microphone connected to a computer in a quiet room. The Greek narrations were recorded in a sound isolating booth of the PhonLab at the University of Konstanz (LingLabs). The video clips were edited and cut with Audacity and stored as .mp3 files. The video clips as well as the narrations were then edited and cut with DaVinci Resolve 17.2 and stored as .mp4 files at a resolution of "720 × 480 NTSC DV."

3.7. Procedure

The tasks were created and hosted on the online platform "Gorilla Experiment Builder" (Anwyl-Irvine et al., 2019). For the clips to run smoothly on "Gorilla," the resolution level was set to low, but this did not affect the high quality of the video clips. Participants performed the tasks at home on their computer or laptop at a convenient time in two sessions that took place 1 week apart. Participants were randomly assigned to one of two groups: Group A was assigned a list of non-verbal video clips in the first session and a list of verbal video clips in the second session; Group B was assigned a list with verbal video clips in the first session and non-verbal video clips in the second session. In the first session, participants were first informed about the content and the procedure of the first session and were asked to check whether they met the criteria for participating in the study. After agreeing to the general conditions, they filled out a short questionnaire about their demographic and linguistic profile. Then they were given short verbal instructions about what they had to do in the experiment: they were asked to watch short video clips (with or without narration, depending on the version) and to decide on the correct answer about what actor1 would logically do after the end of each video clip (in Conditions 1 and 2) and what the observer thinks actor1 would logically do first after the end (in Conditions 3, 4, 5 and 6), by choosing one of the two thought bubbles that were provided. They were asked to make this decision as quickly and as correctly as possible. To familiarize themselves with the task, participants were presented with three practice video clips (all in Condition 1). Participants were able to take a break after half of the video clips. After the end of the task, participants were asked to report any technical or content-related difficulties that they faced during the task. In the second session, participants completed the other version of the task (verbal or non-verbal) but were assigned a different list to ensure that they did not watch exactly the same condition of each scenario. Each session lasted ∼30 min.

3.8. Scoring and analyses

Participants' answer accuracy (correct answer when clicking on the correct thought bubble, incorrect answer when clicking on the incorrect thought bubble) were automatically recorded *via* the platform "Gorilla". The answers indicating the accuracy scores were recorded as correct (1) or incorrect (0).

Since we used a number of different scenarios and lists, it was important to investigate first if participants performed differently in the individual scenarios and lists. To do that we first conducted descriptive statistics to compare scenarios and lists with each other (see sections 4.2.1, 4.2.2, 5.2.1, and 5.2.2). These analyses were followed by a series of generalized linear mixed models to address effects of the data variables we manipulated in the task, i.e., order (first/second), belief (TB/FB), scenario types (motivation and type of change), presentation type (verbal/non-verbal), as well as differences between the order of presentation of non-verbal and the verbal task, i.e., group (group A = non-verbal first/group B = verbal first; see section 4.2.3 and 5.2.3).

The data from the German and Greek participants were analyzed separately in Study 1 (German data) and Study 2 (Greek data).

4. Study 1: German data

4.1. Participants

Fifty neurotypical native German-speaking adults (mean age = 27.64, SD = 4.32; age range: 19–35; 25 male, 25 female) participated in Study 1. Participants were naive to the aim of the experiment. All participants grew up with German as the only language before the age of six and reported no history of language or cognitive delays/impairments, and no history of autism diagnosis. Participants were recruited through personal and academic contacts, through the university's participant recruitment platform "SONA" and through Prolific. Participants received monetary compensation for their participation (10 Euros), with the exception of 22 participants recruited through personal contacts. The research was approved by the Research Ethics Committee of the University of Konstanz (Ethics approval number: 05/2021). Participants gave written informed consent before taking part in the study. All participants had an average correct answer rate of 60% or higher.

4.2. Results

4.2.1. Results for individual scenarios

The proportion of correct answers to each individual scenario is depicted in Figure 1, with numbers given in Supplementary Table 1.[2]

2 Note that since each participant saw only one instance of each scenario, there are no mean values or standard deviations across participants for scenarios. We are providing the proportion of correct answers for each scenario instead as an indication of scenario difficulty.

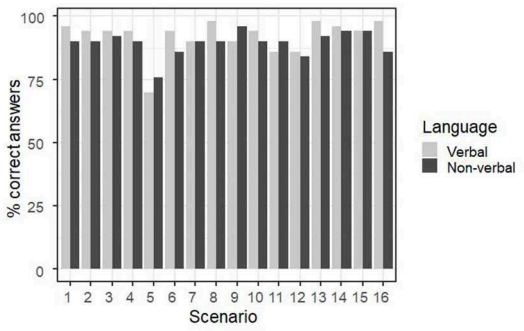

FIGURE 1
German data, proportion of participants who gave accurate answers for each scenario. Proportions of accurate answers for the verbal and non-verbal versions of the task are depicted side by side.

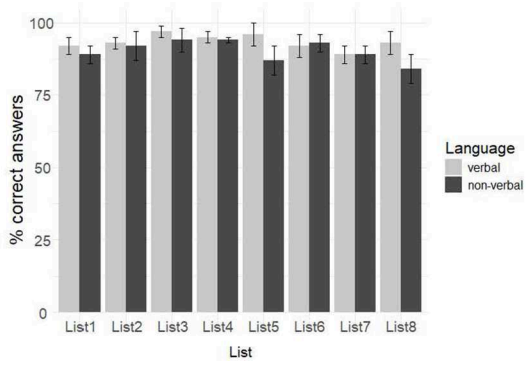

FIGURE 2
German data, mean accuracy rates for each list across participants. Accuracy rates for the verbal and non-verbal versions of each list are depicted side by side.

Performance was close to ceiling, with most scenarios receiving correct responses in above 80% of the cases. This was expected, since the task was meant to not pose any difficulty to neurotypical adults. The only exception was scenario 5, which had a mean accuracy rate of 70% for the verbal version, and 76% for the non-verbal version. Since the mean accuracy rate for this scenario was below 75%, we removed this scenario from all further analysis and also removed it from the toolkit. Both verbal and nonverbal versions had high reliability. Cronbach's alpha for the scenarios in the four experimental conditions (first-order TB, first-order FB, second-order TB, and second-order FB) was for the verbal version 0.811, for the non-verbal version 0.883, and for the combined verbal and non-verbal version 0.845.

4.2.2. Results for individual lists

Mean correct answer rates across participants for each list are depicted in Figure 2, with numbers given in Supplementary Table 2. Mean correct answer rates were well above 80%, indicating that no particular list was remarkably more difficult than the others.

4.2.3. Effects of the variables manipulated

In the following, we describe and monitor the influence of different manipulated variables on participant responses in the critical conditions (excluding the fillers for statistical analysis), using a series of generalized linear mixed models. Models were built using the glmer function of the lme4 package (Bates et al., 2015) in R (Core Team, 2021; family set to binomial, nAQG set to 0). We ran three different analyses exploring the effects of condition, scenario and group on accuracy of performance.

In the first analysis, we monitored the influence of different conditions on answer accuracy. For this, we defined the factors ORDER (first/second) and BELIEF (TB/FB) to address the influence of different conditions, and added the factor PRESENTATION MODE (verbal/non-verbal). In the second analysis, we monitored the influence of different scenario types on answer accuracy; we defined the factors MOTIVATION.CHANGE (deception/no deception) and OUTCOME.CHANGE (unexpected content/unexpected location), again adding PRESENTATION MODE to the factors of the analysis. In the third analysis, we monitored the influence of the participant group, i.e., whether the verbal or non-verbal version of the task was presented first. For this analysis, we defined the factors GROUP (group A = non-verbal first/group B = verbal first) and PRESENTATION MODE.

Since one of the research questions of the current study is to investigate the extent to which the modality of the task (verbal vs. non-verbal) affected performance accuracy, the factor PRESENTATION MODE was included in each model. Interactions were resolved hierarchically. Specifications for individual models are given below. Packages used for data preparation, analysis and plotting are given in the Supplementary Appendix. We only report statistically significant effects unless explicitly stated otherwise.

4.2.3.1. Accuracy rates for conditions; the influence of order and belief

A graphic depiction of mean rates of correct answers per condition is given in Figure 3. An overview of the mean correct answer rates per condition is given in Supplementary Table 3.

Accuracy measures in the verbal and non-verbal version of the task were high across all conditions, with the lowest mean accuracy rate being at 65% (2nd order FB, non-verbal version). Descriptively, accuracy rates were lower for FB than TB conditions; this applied to the non-verbal task for both first- and second-order conditions, and to the verbal task for 2nd order conditions only. The drop in answer accuracy for FB relative to

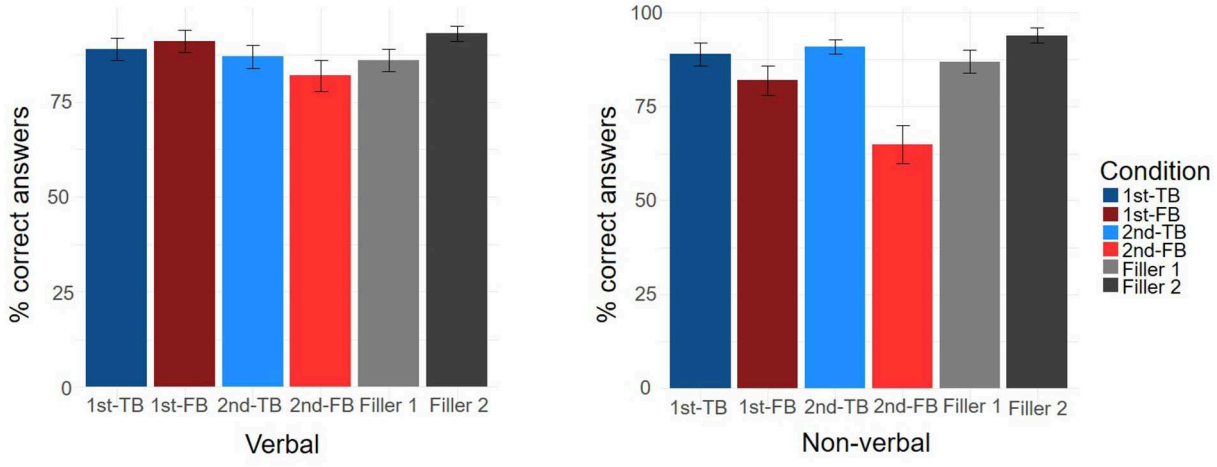

FIGURE 3
German data, mean percentage of correct answers per condition across participants. The left panel shows data for the verbal version, the right panel shows data for the non-verbal version. Error bars depict the standard error of the mean across participants.

TB conditions was more pronounced in the non-verbal than in the verbal version.

To assess how accuracy rates were influenced by condition, and more specifically by truth, order of beliefs, and the version of the task, we analyzed answers using Model 1 (fixed effects for this model and follow-up models are in the Supplementary Table A.1). We specified the main effects and full interactions of ORDER, BELIEF and PRESENTATION MODE as fixed effects, and participant and scenario as random intercepts. PRESENTATION MODE was specified as a random slope for both participant and scenario; for participant, additional random slopes were the main effects and interactions of ORDER and BELIEF. There were main effects of ORDER ($z = 2.46$, $p < 0.05$) and PRESENTATION MODE ($z = -2.34$, $p < 0.05$), showing overall higher accuracy for first- compared to second-order and for the verbal compared to the non-verbal version. There was also an interaction of ORDER and BELIEF ($z = -2.27$, $p < 0.05$). To resolve the interaction of ORDER and BELIEF, we pursued the main effect of BELIEF separately for both levels of ORDER (Model 1a), and the main effect of ORDER separately for both levels of BELIEF (Model 1b). The simple main effect of BELIEF was significant for second-order conditions ($z = -2.73$, $p < 0.01$), showing higher accuracy in TB compared to FB but not for first-order conditions ($p > 0.7$). The simple main effect of ORDER was significant for FB conditions ($z = 2.28$, $p < 0.05$), showing higher accuracy in first- compared to second-order but not for TB conditions ($p > 0.4$). While the lower answer accuracy for non-verbal relative to verbal versions of the task led to a significant influence of the factor PRESENTATION MODE, interactions of PRESENTATION MODE with the other factors did not reach statistical significance. This suggests that the non-verbal version of the task was generally more challenging for the participants than the verbal version across conditions.

4.2.3.2. Accuracy rates for scenario types; the influence of motivation and type of change

In the stimulus list, we carefully balanced scenario types, i.e., the motivation and type of change depicted in the videos. A graphic depiction of mean rates of correct answers per scenario type is illustrated in Figure 4. An overview of the mean correct answer rates per scenario type is given in Supplementary Table 4.

The influence of the factors MOTIVATION.CHANGE, OUTCOME.CHANGE and PRESENTATION MODE was analyzed in Model 2. We specified the main effects and full interactions of MOTIVATION.CHANGE, OUTCOME.CHANGE and PRESENTATION MODE as fixed effects, and participant and scenario as random intercept. For participant, the main effects and interaction of MOTIVATION.CHANGE and OUTCOME.CHANGE and the main effect of PRESENTATION MODE were specified as random slope. For scenario, PRESENTATION MODE was specified as random slope. An overview of the fixed effects for Model 2 is given in the Supplementary Table A.1. There was a marginally significant main effect of PRESENTATION MODE ($z = -1.91$, $p < 0.06$), since rates of correct answers were slightly higher in the verbal (mean across scenario types = 93%) than in the non-verbal version (mean across scenario types = 90%). No main effects or interactions of MOTIVATION.CHANGE or OUTCOME.CHANGE were found, and no significant differences emerged between scenario types, indicating that participants' performance was not affected by these factors.

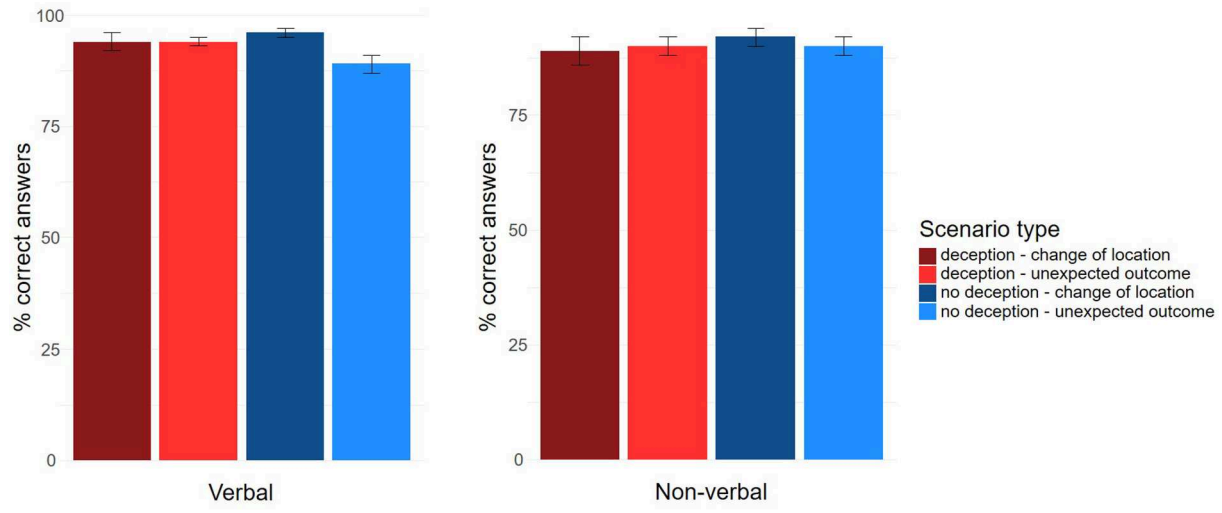

FIGURE 4
German data, mean percentage of correct answers per scenario type across participants. The left panel shows data for the verbal version, the right panel shows data for the non-verbal version. Error bars depict the standard error of the mean across participants.

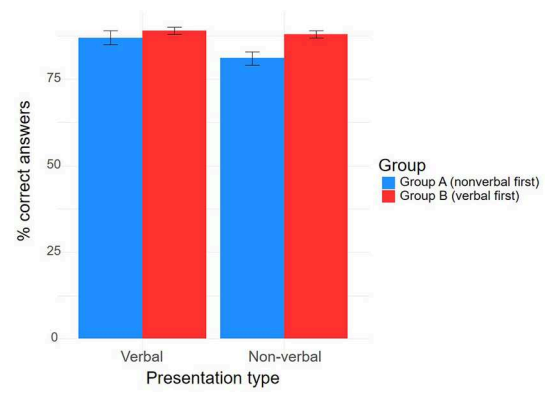

FIGURE 5
German data, mean percentage of correct answers per group across participants, verbal and non-verbal version. Error bars depict the standard error of the mean across participants.

4.2.3.3. Accuracy rates for participant groups; the influence of the presentation order of task versions

Each participant was tested on both the verbal and the non-verbal version of the task. Half of the participants (Group A) were tested on the non-verbal version first, the rest (Group B) on the verbal version first. An illustration of mean rates of correct answers per scenario type is provided in Figure 5. An overview of the mean correct answer rates per group for both versions of the task is provided in Supplementary Table 5.

The influence of the factors GROUP and PRESENTATION MODE was analyzed in Model 3. We specified the main effects and interaction of GROUP and PRESENTATION MODE as fixed effects, and participant and scenario as random intercepts. The main effect of PRESENTATION MODE was specified as random slope for participant, and the main effects and interaction of GROUP and PRESENTATION MODE were specified as random slope for scenario. A full table of the fixed effects for the model is given in the Supplementary Table A.1. There were statistically significant main effects of GROUP ($z = -2.64$, $p < 0.01$) and PRESENTATION MODE ($z = -2.65$, $p < 0.01$). Correct answer rates were higher in the verbal (mean = 88%) than in the non-verbal version (mean = 85%), and higher for Group B (verbal first, mean = 89%) than for Group A (non-verbal first, mean = 84%).

4.3. Summary of results

In one of 16 scenarios (scenario 5), participants consistently performed worse than in the others. This scenario was excluded from all subsequent data analyses and will not be included in the final version of the toolkit. Performance did not differ across lists, indicating that the remaining scenarios were similar in difficulty, even when presented in different combinations and conditions. A comparison between the different conditions revealed that for second-order beliefs, FBs were more difficult than TBs. For first-order beliefs, the difference between FBs and TBs did not reach significance. Performance in the non-verbal task was worse than in the verbal task; however, it was considerably above chance even in the most difficult conditions (second-order FB). Descriptively, the differences between conditions had the same direction, but were more pronounced in the non-verbal than in the verbal task. There was

no influence of scenario type, indicating that for neurotypical adults, scenarios without deception were not any more difficult than scenarios with deception. There was an influence of the order in which the two versions (i.e., verbal vs. non-verbal) of the task were shown. Participants who had seen the verbal version first performed better than participants who had seen the non-verbal version first. While accuracy rates in both groups were high and the absolute difference was not dramatic (5%), the difference still reached statistical significance.

5. Study 2: Greek data

5.1. Participants

Fifty neurotypical native Greek-speaking adults (mean age = 27.34, SD = 5.12; age range: 18–35; 25 male, 25 female) participated in Study 2. Participants were naive to the aim of the experiment. All participants grew up with Greek as the only language before the age of six and reported no history of language or cognitive delays/impairments, and no history of autism diagnosis. Participants were recruited through personal contacts and through Prolific. Participants received monetary compensation for their participation (10 Euros), with the exception of 22 participants recruited through personal contacts. Participants gave written informed consent before taking part in the study. All participants had an average correct answer rate of 60% or higher.

5.2. Results

The presentation and analysis of results is parallel to the one for the German data in section 4.

5.2.1. Results for individual scenarios

The proportion of participants who gave accurate answers to individual scenarios is depicted in Figure 6, with numbers provided in Supplementary Table 6.

Performance was close to ceiling, with most scenarios receiving correct responses in over 80% of the cases, which verifies our hypothesis that the task was not difficult for neurotypical adults. Scenario 5 seemed to be the only exception, with an accuracy rate of 80% for the verbal version, and 76% for the non-verbal version, mirroring the German results. Due to the relatively high difficulty associated with Scenario 5, and in order to keep the analysis parallel to the one for the German dataset, we removed it from the dataset before any further analysis. Both verbal and nonverbal versions had high reliability. Cronbach's alpha for the scenarios in the four experimental conditions (first-order TB, first-order FB, second-order TB, and

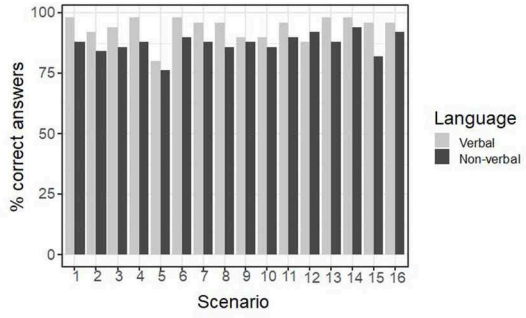

FIGURE 6
Greek data, proportion of participants who gave accurate answers for each scenario. Proportions of accurate answers for the verbal and non-verbal versions of the task are depicted side by side.

second-order FB) was for the verbal version 0.881, for the non-verbal version 0.868, and for the combined verbal and non-verbal version 0.908.

5.2.2. Results for individual lists

Mean correct answer rates across participants for each list are illustrated in Figure 7, with numbers provided in Supplementary Table 7. Mean correct answer rates in general were well above 80%, indicating that no particular list was more difficult than the others. The only exception was List 1, where the non-verbal version had a relatively low mean answer rate of 75%. This was mainly due to one participant with a mean answer rate below 50% in the non-verbal version; this participant scored at 87% in the verbal version.

5.2.3. Effects of the variables manipulated

In the following section, we describe and monitor the influence of different manipulated variables on participant responses in the critical conditions. Analyses were performed in parallel to the German version.

5.2.3.1. Accuracy rates for conditions; the influence of order and belief

Mean rates of correct answers per condition are illustrated in Figure 8, for the verbal and the non-verbal version respectively. An overview of the mean correct answer rates per condition is provided in Supplementary Table 8.

Answer accuracies in the verbal and non-verbal version of the task were high across all conditions, with the lowest mean accuracy rate being at 73% for the second-order FB trials of the non-verbal version. Accuracy rates were lower for FB than TB conditions; this applied to the verbal task for both first-

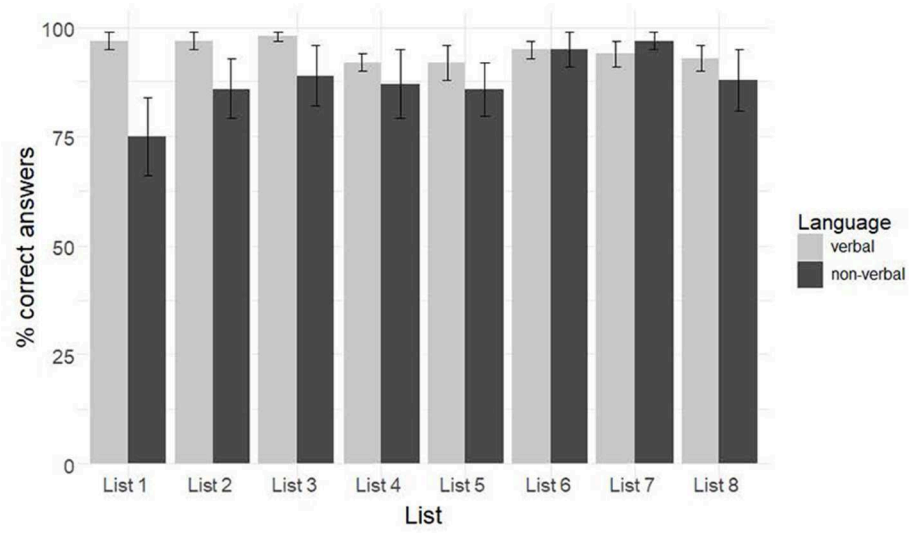

FIGURE 7
Greek data, mean accuracy rates for each list across participants. Accuracy rates for the verbal and non-verbal versions of each list are depicted side by side.

and second-order conditions, and to the non-verbal task for the second-order condition.

To assess how accuracy rates were influenced by condition, and more specifically by truth and order of beliefs, and the version of the task, we analyzed answers using Model 1 (see the preceding section). The random slopes for scenario included main effects and interaction of ORDER and BELIEF, in addition to the main effect of PRESENTATION MODE. Fixed effects for this model are in the Supplementary Table A.2. There was a main effect of ORDER ($z = -5.39$. $p < 0.05$). No other main effects or interactions reached significance.

5.2.3.2. Accuracy rates for scenario types; the influence of motivation and type of change

Mean rates of correct answers per scenario type are illustrated in Figure 9. An overview of the mean correct answer rates per scenario type is provided in Supplementary Table 9.

The influence of the factors MOTIVATION.CHANGE, OUTCOME.CHANGE and PRESENTATION MODE was analyzed in Model 2. A fixed effects table is provided in the Supplementary Table A.2.

There was a marginally significant effect of PRESENTATION MODE ($z = -1.90$, $p < 0.06$), with no other main effects or interactions. Accuracy rates were slightly lower in the non-verbal (mean = 88%) than in the verbal version (mean = 95%), with no significant difference between scenario types. No main effects or interactions of MOTIVATION.CHANGE or OUTCOME.CHANGE were found; differences between scenario types were also not significant, which suggests that participants' performance was not affected by these factors.

5.2.3.3. Accuracy rates for participant groups; the influence of the presentation order of task versions

Each participant was tested on both the verbal and the non-verbal version of the task. Half of the participants (Group A) were tested on the non-verbal version first, and the rest (Group B) were tested on the verbal version first. Mean rates of correct answers per scenario type are illustrated in Figure 10. An overview of the mean correct answer rates per group for both versions of the task is provided in Supplementary Table 10.

The influence of the factors GROUP and PRESENTATION MODE was analyzed in Model 3. A full table of the fixed effects for the model is given in the Supplementary Table A.2. There were statistically significant main effects of GROUP ($z = -2.07$, $p < 0.05$) and PRESENTATION MODE ($z = -2.88$, $p < 0.01$). Correct answer rates were higher in the verbal (mean = 89%) than in the non-verbal version (mean = 83.5%), and higher for Group B (verbal first, mean = 89%) than for group A (non-verbal first, mean = 83.5%).

5.3. Summary of results

In one of the 16 scenarios, participants performed worse than in the others. Since this scenario had also been more error-prone in the German version of the task, we excluded it from all subsequent data analyses. Performance across the different

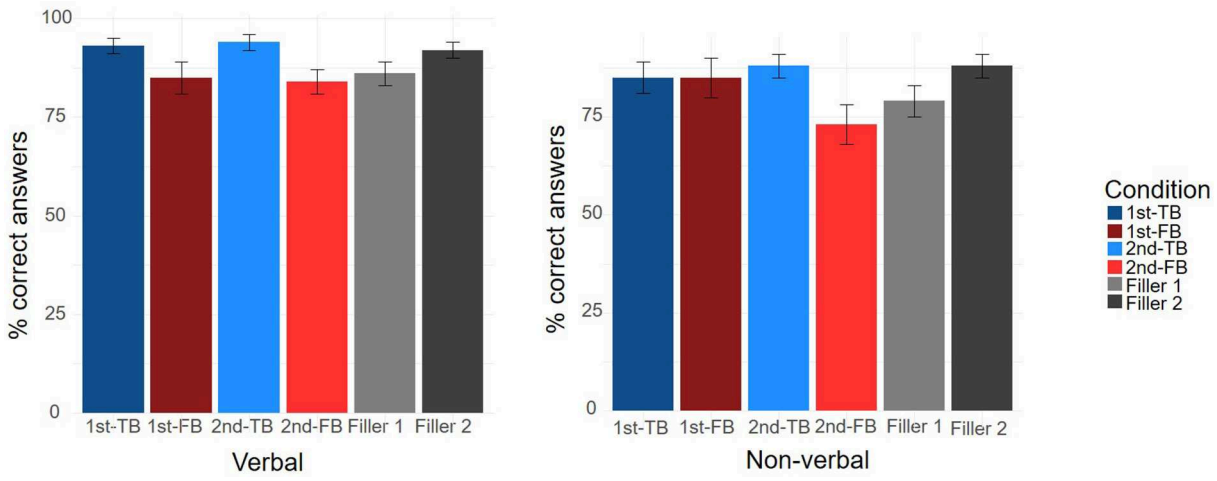

FIGURE 8
Greek data, mean percentage of correct answers per condition across participants. The left panel shows data for the verbal version, the right panel shows data for the non-verbal version. Error bars depict the standard error of the mean across participants.

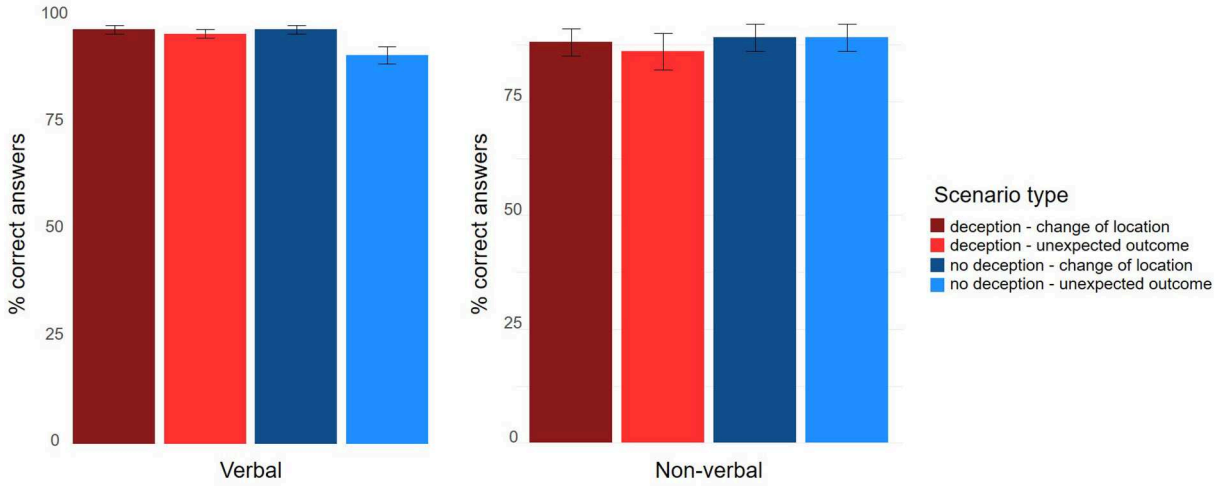

FIGURE 9
Greek data, mean percentage of correct answers per scenario type across participants. The left panel shows data from the verbal version, the right panel shows data from the non-verbal version. Error bars depict the standard error of the mean across participants.

lists did not differ, indicating that the remaining scenarios were similar in difficulty, even when presented in different combinations and conditions (an exception for the non-verbal version of List 1 seems to be driven by a single participant who found the non-verbal version of the task particularly difficult).

A comparison across the different conditions revealed that second-order beliefs were more difficult than first-order beliefs. Descriptively, FBs were more difficult than TBs in the verbal version, and the same pattern held for second-order beliefs only in the non-verbal version; however, no main effects or interactions of BELIEF or PRESENTATION MODE reached significance. In general, performance in all conditions was high and remained well above chance even in the most difficult condition (second-order FB). Descriptively, the differences between conditions had the same direction, but were more pronounced in the verbal than in the non-verbal task. There was no influence of scenario type, indicating that for neurotypical adults, scenarios involving deception were not any more difficult

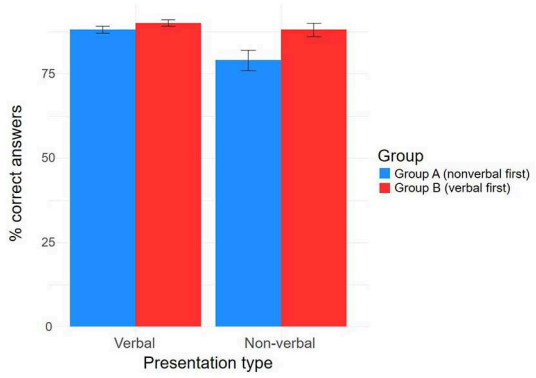

FIGURE 10
Greek data, mean percentage of correct answers per group across participants, verbal and non-verbal version. Error bars depict the standard error of the mean across participants.

than scenarios without deception. There was an influence of the order in which the versions of the task were shown. Specifically, as in Study 1, participants who had seen the verbal version first performed better than participants who had seen the non-verbal version first. While performance rates in both groups were high and the absolute difference was not dramatic (5.5%), the difference still reached statistical significance.

6. Discussion

ToM is related to the development of language in a bidirectional way; the emergence of ToM enables language development (Nelson, 2005), whereas language skills promote ToM abilities (Harris et al., 2005; Milligan et al., 2007). Moreover, language skills and in particular syntax have been argued to be important in the attribution of beliefs, the development of ToM, and the success in ToM tasks (de Villiers and de Villiers, 2009, 2014). ToM is a vulnerable area in a number of populations. For example, autistic individuals have been shown to have difficulties in performing successfully in FB tasks that involve first- and second-order ToM with unexpected change of contents (e.g., the Smarties task) and unexpected change of location (e.g., the Sally-Anne task).

A large proportion of autistic individuals have low or minimal language skills (e.g., Lord and Paul, 1997; Lord et al., 2004). And yet, the most frequently used ToM tasks put high demands on language because they include advanced vocabulary, they involve long narrations, and require participants to respond to complex questions (e.g., Norbury and Bishop, 2003; Durrleman et al., 2016; Peristeri et al., 2017). In other words, the most frequently used ToM tasks confound ToM with language. This makes it difficult to disentangle whether low performance in ToM tasks is due to the individuals' low language or to their mentalizing abilities. Therefore, there is a need to develop ToM tasks that exert low demands on language. The only studies to date that have employed low-verbal ToM tasks with children either included only first-order ToM (Durrleman et al., 2016; Peristeri et al., 2021) or included both first- and second-order ToM tasks but required responses to complex questions (Hollebrandse et al., 2014). Finally, the most widely used ToM tasks have employed static images or puppets and neglected the complex and dynamic character of the real world, which requires extracting information from incoming input, developing predictions about upcoming events, and updating these predictions based on people's behavior. Therefore, it is unclear whether successful performance in such tasks reflects the ability to use ToM in real-life situations.

The current research aimed at addressing these limitations by developing a comprehensive toolkit that tests both first- and second-order ToM in a task that puts low demands on language in terms of the instructions and the individuals' responses, and includes both a verbal and a non-verbal modality. This makes the task appropriate for both individuals with low and high verbal abilities and can assess ToM mediated by language as well as ToM independent of language. Moreover, the task is based on interactive, real-life scenarios, using video stimuli that require developing and updating predictions based on the behavior of real-life people and situations. Therefore, performance in the task is more likely to reflect the ability to use ToM in real-life.

Our ToM toolkit included a large number of items divided across several types of scenarios (change of location/unexpected content and deception/non-deception) to address how individuals react to different scenario types and to be able to exclude scenarios that would be problematic for large numbers of participants (adults and/or children) at a later stage in the task development. Finally, given the shortage in ToM tasks and data from languages other than English, we developed the toolkit in two languages, German and Greek, in order to make it accessible to a large number of individuals across several countries as well as to multilingual speakers. This makes the toolkit appropriate for at least two European societies and it can be easily adapted to others.

Neurotypical adults performed this task in the verbal and non-verbal modalities in two distinct studies, one in Germany and one in Greece. The two studies produced very similar results. Results were comparable for all scenarios, with the exception of a single scenario (scenario number 5) which produced low accuracy in both studies, and was removed from the data analysis. Neurotypical adults performed very well, but not completely at ceiling, especially in conditions which require more complex FB reasoning. Descriptively, the percentage of correct answers was lower for conditions requiring the use of second-order beliefs than for those requiring the

use of first-order beliefs. Likewise, the proportion of correct answers was lower for conditions requiring comprehension of FBs than for those requiring the comprehension of TBs. We take this to reflect the increased difficulty for conditions with false and second-order beliefs (see also Bernstein et al., 2017).

The motivation (deception, no deception) and result of the depicted change (unexpected change of location, unexpected content) did not affect performance in the different versions and languages of the task, showing that these factors do not influence task difficulty for neurotypical adults. This may be different for young children, autistic individuals, and other populations whose ToM skills may be impaired. The present findings provide important data that can be compared against data from the aforementioned populations. Should these future studies find an influence of motivation and change of result on performance for different participant groups, this could be interpreted as a true reflection of difficulties, rather than a task-based confound.

Generally, performance in both verbal- and non-verbal versions of the task was high. However, accuracy rates were lower in the non-verbal than in the verbal version of the task, as in the Hollebrandse et al. (2014) task with neurotypical children. Beginning with the verbal task had a visible training effect, leading to significantly better performance in the non-verbal task for participants in the verbal-first group in contrast to the group that was administered the non-verbal task first. These findings suggest that neurotypical adults use language to mediate ToM performance and learn from a language-mediated task when performing, at a later stage, a non-verbal ToM task. This information is important and should be taken into account when considering further applications of the toolkit with children and adults with low language skills both in experimental follow-up studies and when interpreting the results of the toolkit in a diagnostic setting. If both versions of the task are used, they should be presented in the same order, with the non-verbal version presented first to make sure that there is no influence from the verbal to the non-verbal task.

While the general tendencies of results were the same for both studies and languages, there were also subtle differences. In German, the difference between TBs and FBs was more pronounced for second- (vs. first-) order. This was not replicated in Greek; here, the increased difficulty with higher orders surfaced as a simple main effect of order, not as an interaction with belief. We assume that these differences do not reflect language-specific processing strategies, but rather the subtle shifts in accuracy rates resulting from random differences between groups. Even though all individual participants in our studies showed good general performance, tendencies for individual participants in each group may have affected the significance of two- or three-way interactions.

7. Conclusions

The aim of this research was to fill the need of assessing first- and second-order ToM in both verbal and non-verbal modalities that would be appropriate for individuals with high and low verbal abilities. Toward that aim, we developed a comprehensive toolkit that tests both first- and second-order ToM *via* FB tasks in the non-verbal and verbal modality using interactive, real-life scenarios. The toolkit was developed for German and Greek and was balanced for factors that may influence performance, i.e., deception, change of location and unexpected content. To validate the toolkit, two studies were conducted with adult neurotypical individuals in Germany and Greece. The data from the two studies show similar patterns of results. There was high performance in all conditions with no effects of scenario, deception or type of outcome. However, second-order conditions were slightly more challenging than first-order conditions, and the non-verbal version was also more challenging than the verbal version of the task. Participants being first administered the verbal version and then the non-verbal version performed slightly better than those taking the opposite order, suggesting that neurotypical adults use language to mediate ToM performance and learn from a language-mediated task when performing, a non-verbal ToM task. The results suggest that the toolkit is suitable for neurotypical adults, and is expected to be well inside the sensitive range for children and autistic individuals. The toolkit can be easily adapted to other languages, and scenarios can be combined freely to meet the needs of future research with neurotypical children, autistic individuals with both high and low verbal abilities, as well as other populations that have been shown to have difficulties in ToM. Importantly, the toolkit can assess ToM mediated by language, as well as ToM independent of language.

Data availability statement

The datasets presented in this article are not readily available because the analyzed datasets for this study will be uploaded in the KonDATA depository of the University of Konstanz. Requests to access the datasets should be directed at: https://kondata.uni-konstanz.de/index_en.html.

Ethics statement

The studies involving human participants were reviewed and approved by University of Konstanz Research Ethics Committee. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individuals for the publication of any potentially identifiable images or data included in this article.

Author contributions

FB created the video clips. DB, AC, and AG analyzed the data. MA, AC, TM, EP, and AT wrote the manuscript. All authors contributed to the design of the toolkit. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the Federal Ministry of Education and Research, Germany and the State Scholarships Foundation, Greece, as part of the IKYDA Project German-Greek toolkit for Theory of Mind and Language in Autism (MiLA), Project-ID: 57547309, Principal Investigator in Germany: TM; Principal Investigator in Greece: AT. The responsibility for the content of this publication lies with the authors.

Acknowledgments

We would like to thank the participants of the study for their willingness to take part.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/flang.2022.1052095/full#supplementary-material

References

Andreou, M., and Skrimpa, V. (2020). Theory of mind deficits and neurophysiological operations in autism spectrum disorders: a review. *Brain Sci.* 10, 393. doi: 10.3390/brainsci10060393

Andreou, M., and Skrimpa, V. (2022). Re-examining labels in neurocognitive research: evidence from bilingualism and autism as spectrum-trait cases. *Brain Sci.* 12, 1113. doi: 10.3390/brainsci12081113

Andreou, M., Tsimpli, I. M., Durrleman, S., and Peristeri, E. (2020). Theory of mind, executive functions, and syntax in bilingual children with autism spectrum disorder. *Languages* 5, 67. doi: 10.3390/languages5040067

Anwyl-Irvine, A. L., Massonié, J., Flitton, A., Kirkham, N. Z., and Evershed, J. K. (2019). Gorilla in our midst: an online behavioural experiment builder. *Behav. Res. Methods.* 52, 388–407. doi: 10.1101/438242

Arslan, B., Hohenberger, A., and Verbrugge, R. (2017). Syntactic recursion facilitates and working memory predicts recursive theory of mind. *PLoS ONE* 12, e0169510. doi: 10.1371/journal.pone.0169510

Balaban, N., Friedmann, N., and Ziv, M. (2016). Theory of mind impairment after right-hemisphere damage. *Aphasiology* 30, 1399–1423. doi: 10.1080/02687038.2015.1137275

Baldimtsi, E., Nicolopoulou, A., and Tsimpli, I. M. (2021). Cognitive and affective aspects of theory of mind in Greek-speaking children with autism spectrum disorders. *J. Autism Dev. Disord.* 51, 1142–1156. doi: 10.1007/s10803-020-04595-0

Barnes, J. L., and Baron-Cohen, S. (2012). The big picture: storytelling ability in adults with autism spectrum conditions. *J. Autism Dev. Disord.* 42, 1557–1565. doi: 10.1007/s10803-011-1388-5

Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/4635.001.0001

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., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arXiv [Preprint]*. arXiv: 1506.04967. Available online at: https://arxiv.org/pdf/1506.04967.pdf

Begeer, S., Gevers, C., Clifford, P., Verhoeve, M., Kat, K., Hoddenbach, E., et al. (2011). Theory of mind training in children with autism: a randomized controlled trial. *J. Autism Dev. Disord.* 41, 997–1006. doi: 10.1007/s10803-010-1121-9

Bernstein, D. M., Coolin, A., Fischer, A. L., Thornton, W. L., and Sommerville, J. A. (2017). False-belief reasoning from 3 to 92 years of age. *PLoS ONE* 12, e0185345. doi: 10.1371/journal.pone.0185345

Bodner, K. E., Engelhardt, C. R., Minshew, N. J., and Williams, D. L. (2015). Making inferences: comprehension of physical causality, intentionality, and emotions in discourse by high-functioning older children, adolescents, and adults with autism. *J. Autism Dev. Disord.* 45, 2721–2733. doi: 10.1007/s10803-015-2436-3

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

Chevallier, C., Parish-Morris, J., Tonge, N., Le, L., Miller, J., Schultz, R. T., et al. (2014). Susceptibility to the audience effect explains performance gap between children with and without autism in a theory of mind task. *J. Exp. Psychol. Gen.* 143, 972–979. doi: 10.1037/a0035483

Chevallier, C., Wilson, D., Happé, F., and Noveck, I. (2010). Scalar inferences in autism spectrum disorders. *J. Autism Dev. Disord.* 40, 1104–1117. doi: 10.1007/s10803-010-0960-8

Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://R-project.org/

de Villiers, J. G., and de Villiers, P. A. (2009). "Complements enable representation of the contents of false beliefs: the evolution of a theory of theory of mind," in *Language Acquisition. Palgrave Advances in Linguistics*, ed S. Foster-Cohen (London: Palgrave Macmillan), p. 169–195. doi: 10.1057/9780230240780_8

de Villiers, J. G., and de Villiers, P. A. (2014). The role of language in theory of mind development. *Top. Lang. Disord.* 34, 313–328. doi: 10.1097/TLD.0000000000000037

Diehl, J. J., Friedberg, C., Paul, R., and Snedeker, J. (2015). The use of prosody during syntactic processing in children and adolescents with autism spectrum disorders. *Dev. Psychopathol.* 27, 867–884. doi: 10.1017/S0954579414000741

Durrleman, S., Burnell, M., Thommen, E., Foudon, N., Sonie, S., Reboul, A., et al. (2016). The language cognition interface in ASD: complement sentences and false belief reasoning. *Res. Autism Spectr. Disord.* 21, 109–120. doi: 10.1016/j.rasd.2015.10.003

Durrleman, S., and Delage, H. (2020). Training complements for belief reasoning in developmental language disorder. *J. Speech Lang. Hear. Res.* 63, 1861–1877. doi: 10.1044/2020_JSLHR-19-00075

Durrleman, S., Peristeri, E., and Tsimpli, I. M. (2022a). The language-communication divide: evidence from bilingual children with atypical development. *Evol. Linguist. Theory* 4, 5–53. doi: 10.1075/elt.00037.dur

Durrleman, S., Tsimpli, I. M., and Peristeri, E. (2022b). "Domino effects of bilingualism in autism spectrum disorders? Executive functions, complement clauses and theory of mind," in *BUCLD 46: Proceedings of the 44th Annual Boston University Conference on Language Development*, eds Y. Gong, and F. Kpogo, Vol. 1 (Somerville, MA: Cascadilla Press), 180–193.

Flobbe, L., Verbrugge, R., Hendriks, P., and Krämer, I. (2008). Children's application of theory of mind in reasoning and language. *Log. Lang. Inf.* 17, 417–442. doi: 10.1007/s10849-008-9064-7

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

Frith, C. D., and Corcoran, R. (1996). Exploring 'theory of mind' in people with schizophrenia. *Psychol. Med.* 26, 521–530. doi: 10.1017/S0033291700035601

Happé, F. G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Dev.* 66, 843–855. doi: 10.2307/1131954

Harris, P. L., de Rosnay, M., and Pons, F. (2005). Language and children's understanding of mental states. *Curr. Dir. Psychol. Sci.* 14, 69–73. doi: 10.1111/j.0963-7214.2005.00337.x

Heavey, L., Phillips, W., Baron-Cohen, S., and Rutter, M. (2000). The awkward moments test: a naturalistic measure of social understanding in autism. *J. Autism Dev. Disord.* 3, 225–236. doi: 10.1023/A:1005544518785

Hollebrandse, B., van Hout, A., and Hendriks, P. (2014). Children's first and second-order false-belief reasoning in a verbal and a low-verbal task. *Synthese* 191, 321–333. doi: 10.1007/s11229-012-0169-9

Korkmaz, B. (2011). Theory of mind and neurodevelopmental disorders of childhood. *Pediatr. Res.* 69, 101–108. doi: 10.1203/PDR.0b013e318212c177

Lecce, S., Caputi, M., and Pagnin, A. (2014). Long-term effect of theory of mind on school achievement: the role of sensitivity to criticism. *Eur. J. Dev. Psychol.* 11, 305–318. doi: 10.1080/17405629.2013.821944

Lönnqvist, L., Loukusa, S., Hurtig, T., Mäkinen, L., Siipo, A., Väyrynen, E., et al. (2017). How young adults with autism spectrum disorder watch and interpret pragmatically complex scenes. *Q. J. Exp. Psychol.* 70, 2331–2346. doi: 10.1080/17470218.2016.1233988

Lord, C., and Paul, R. (1997). "Language and communication in autism," in *Handbook of Autism and Pervasive Developmental Disorders*, eds D. Cohen, and F. Volkmar (New York, NY: Wiley), 195–225.

Lord, C., Risi, S., and Pickles, A. (2004). "Trajectory of language development in autism spectrum disorders," in *DevelopmentalLanguage Disorders*, eds M. L. Rice, and S. F. Warren (Mahwah, NJ: Erlbaum), 7–30.

Milligan, K., Astington, J., and Dack, L. (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

Nelson, K. (2005). "Language pathways into the community of minds," in *Why Language Matters for Theory of Mind*, eds J. W. Astington, and J. A. Baird (Oxford: Oxford University Press), 26–49. doi: 10.1093/acprof:oso/9780195159912.003.0002

Norbury, C. F., and Bishop, D. V. M. (2003). Narrative skills of children with communication impairments. *Int. J. Lang. Commun. Disord.* 38, 287–313. doi: 10.1080/1368231000010813

Peristeri, E., Andreou, M., and Tsimpli, I. M. (2017). Syntactic and story structure complexity in the narratives of high- and low-language ability children with autism spectrum disorder (special issue: investigating grammar in autism spectrum disorders). *Front. Psychol.* 8, 2027. doi: 10.3389/fpsyg.2017.02027

Peristeri, E., Baldimtsi, E., Vogelzang, M., Tsimpli, I. M., and Durrleman, S. (2021). The cognitive benefits of bilingualism in autism spectrum disorder: is theory of mind boosted and by which underlying factors? *Autism Res.* 14, 1695–1709. doi: 10.1002/aur.2542

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 4, 515–526. doi: 10.1017/S0140525X00076512

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

Siegal, M., and Beattie, K. (1991). Where to look first for children's knowledge of false beliefs. *Cognition* 38, 1–12. doi: 10.1016/0010-0277(91)90020-5

Tager-Flusberg, H. (2000). "Language and understanding minds: connections in autism," in *Understanding Other Minds: Perspectives From Developmental Cognitive Neuroscience*, eds S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen (Oxford: Oxford University Press), 124–149.

Tager-Flusberg, H., and Sullivan, K. (1994). A second look at second-order belief attribution in autism. *J. Autism Dev. Disord.* 24, 577–586. doi: 10.1007/BF02172139

Terzi, A., Marinis, T., and Francis, K. K. (2016). The interface of syntax with pragmatics and prosody in children with autism spectrum disorders. *J. Autism Dev. Disord.* 46, 2692–2706. doi: 10.1007/s10803-016-2811-8

Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev.* 72, 655–84. 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

Zelazo, P. D., Jacques, S., Burack, J. A., and Frye, D. (2002). The relation between theory of mind and rule use: evidence from persons with autism-spectrum disorders. *Infant Child Dev.* 11, 171–95. doi: 10.1002/icd.304