

The scopes, limits and developmental foundations of implicit Theory of Mind

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium

an der Universität Hamburg

Fakultät für Psychologie und Bewegungswissenschaft

Institut für Psychologie

vorgelegt von

Sebastian Dörrenberg

Hamburg, 2019

Promotionsprüfungsausschuss

Vorsitzende:	Prof. Dr. rer. nat. Nale Lehmann-Willenbrock
1. Dissertationsgutachter:	Prof. Dr. rer. nat. Ulf Liskowski
2. Dissertationsgutachter:	Prof. Dr. rer. nat. Hannes Rakoczy
1. Disputationsgutachter:	Prof. Dr. rer. nat. Jan Wacker
2. Disputationsgutachterin:	Prof. Dr. phil. Jenny Wagner
Tag der Disputation:	26.03.2019

Acknowledgments

First of all, I would like to thank my supervisors Hannes Rakoczy and Ulf Liszkowski for giving me, someone with a different background, the opportunity to make a scientific contribution to developmental psychology and infant ToM research in particular. This research project originates mainly from their stock of ideas and expertise. I owe them gratitude for sharing their knowledge and experience with me, for inspiring discussions at all times, for their guidance, and, of course, for broadening my horizons.

Thank you to my team at the Developmental Psychology Department at the University of Hamburg: Jessica, Johanna, Mareike, Marianna, Ranjani, Susanna, Wiebke and all students who kept the lab running. In the past three years, I received social, organizational and scientific support from you guys. I enjoyed each lunchtime, tenderly arranged celebration, joint conference, scientific discussion and gossip with you!

Thank you to my fellows Matthias and Wiebke from the neighbor department for always having an open door and good advice for me.

Thank you to the Crossing Project and all involved PhDs and PIs. We had inspiring and constructive workshops, conferences, project meetings and social evenings. I am very grateful to be a part of this cross-disciplinary collaboration. Thanks also to the DFG that funded this project and my PhD position.

Thank you Lisa Wenzel and Marina Proft from the University of Göttingen for your collaboration. It was (and is) a pleasure to plan studies and write papers with you.

Last but not least, I am grateful for the love and patience that my wife Merle and my daughter Ruby gifted me with during stressful times.

Table of contents

Abstract	5
Introduction	6
Classical findings of Theory of Mind research.....	6
False belief tasks	6
Mindreading before age four.....	7
Developing an explicit Theory of Mind	8
New findings with implicit measures	10
Implicit false belief tasks.....	10
Theoretical accounts on implicit ToM	12
Rationale of the current thesis	13
Research questions.....	13
Work program.....	15
Study 1: How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures.....	17
Abstract	17
Introduction.....	17
Robustness, reliability and replicability of implicit ToM tasks.....	18
Convergent validity of implicit ToM tasks.....	19
Rationale of the present study	20
Material and methods.....	22
Participants	22
Design	23
Set-up and Procedure	24
Coding and analyses	27
Results	29
Anticipatory Looking task	29

Anticipation + Outcome task	33
Interaction task	40
Correlations between measures	41
Discussion	43
Reliability of implicit ToM measures.....	43
Convergent validity of implicit ToM measures	46
Conclusion.....	48
Acknowledgements.....	49
Appendix 1. Stimuli & Procedure	49
Anticipatory Looking task	49
Anticipation + Outcome task	50
Interaction task.....	51
Appendix 2. Results.....	52
Analyses on looking time in A + O task including participants with more weighted gaze samples	52
Mirrored vs. un-mirrored videos in the AL task.....	53
Analyses on distance to the eye-tracker.....	53
Offset illumination analyses for the first trial of the AL task	54
RCD in the fourth second of the reaching phase	55
Study 2: The sefo task: A measure of early false belief understanding?	56
Abstract	56
Introduction.....	56
Methods	61
Direct replication	61
Pragmatically modified task.....	63
Results	66
Direct replication	66
Pragmatically modified task.....	68
Discussion	70

Acknowledgement	74
Study 3: Reliability and generalizability of an acted-out false belief task in 3-year-olds.....	75
Abstract	75
Introduction	75
Methods	81
Participants	81
Design and Procedure	81
Results	85
The Duplo tasks.....	85
Standard change-of-location task and relations to Duplo tasks	86
Comparisons between the two labs	87
Discussion	87
Acknowledgment	93
Study 4: What predicts implicit ToM development?	94
Introduction	94
Methods	94
Results	95
Implicit ToM task	95
Correlations with predictor variables	96
Conclusions	96
General discussion	98
Summary and synthesis of findings	98
Reliability of implicit ToM measures.....	98
Convergent validity	102
Developmental determinants.....	103

Limitations and outlook105

Concluding remarks108

References110

Abstract

How does our capacity to ascribe others subjective perspectives on the world, or “Theory of Mind” (ToM), develop? Traditional task using explicit measures produced comprehensive and converging evidence that only from age four children acquire a ToM. However, the last decade provided an impressive body of evidence on implicit ToM, suggesting that language and socialization play a marginal role in understanding others. That is, different non-verbal paradigms were established, such as violation-of expectation, anticipatory looking or interaction paradigms, which suggest that even very young infants ascribe false beliefs to other agents – the litmus test for understanding subjectivity. While each task itself might reflect a conceptual ToM capacity, for each local finding there might be alternative simpler explanations. In addition, we currently do not know how reliable findings on implicit ToM really are. In older children, numerous studies have shown unity and reliability of explicit ToM. For early implicit ToM, comparable systematic studies of reliability and cross-validations of findings are still lacking. While deflationary accounts would predict no unity, and hence unrelated performances, nativist and (to some degree) two-systems accounts would predict full unity and reliability. Further, for explicit ToM, longitudinal studies found language, executive functions, as well as socio-cognitive skills and socio-pragmatic experiences to be valid developmental predictors. In contrast, for implicit ToM, hardly anything is known to date concerning its developmental foundations. The current project aims at filling these gaps by investigating whether early ToM abilities reflect a robust and unitary ToM capacity in systematic cross-sectional studies, as well as the developmental determinants of early ToM in a longitudinal study.

Introduction

Classical findings of Theory of Mind research

As adult *Homo sapiens*, we understand our fellows as rational agents that possess an inner life and act on subjective mental states. This capacity, which is also called Theory of Mind (ToM; Premack & Woodruff, 1978), enables us to predict and explain the behavior of others and to engage in sophisticated forms of communication and cooperation (e.g., Tomasello, 1999; Tomasello & Rakoczy, 2003). When and how we develop a ToM during ontogeny is a hot topic of developmental psychology since the past 40 years. The classical finding is that children acquire a full-blown concept of others' mental states (meta-representations, or ToM) around age four to five (Wellman, Cross, & Watson, 2001). This ability is reflected in ascribing propositional attitudes, such as desires and beliefs, to another person.

False belief tasks

To demonstrate ToM in a subject, tests are needed that require the understanding that another individual has a subjective representation, or rather misrepresentation, of the real state of affairs (Bennett, 1978; Dennett, 1978). In the classical change-of-location or false belief (FB) task (Wimmer & Perner, 1983), which has become the litmus test for crediting a ToM, participants are told a story of the protagonist Maxi. Maxi puts his chocolate into a green cupboard in the kitchen and leaves the house to go to the playground. In his absence, his mother enters the kitchen, transfers the chocolate from the green to the blue cupboard and leaves. Then Maxi comes back home and wants to eat his chocolate. Participants are asked, "Where will Maxi look for his chocolate?" If participants have a representation of Maxi's FB, then they will answer that he will search in the green cupboard, although they themselves represent the chocolate in the blue cupboard (where it actually is). Children at three years of age typically fail this task by predicting that Maxi will look in the chocolate's actual location, and only from four years on, children systematically pass (see for a meta-analysis Wellman et al., 2001).

Over the years, several superficially different tasks were established to measure the same competence of ascribing FBs to other agents. For instance, in the unexpected-content task (Hogrefe, Wimmer, & Perner, 1986), an experimenter shows a box of stereotypical content (e.g., a box of Smarties) to the child. After opening the box, they figure out that the box actually

contains atypical items (e.g., pencils). After realizing their own initial FB, participants are asked what another child, which was not in the room, would expect to be in that box. Another sort of FB tasks, so called intensionality (or aspectuality) tasks, test children's understanding that an agent represents reality always under specific aspects (Apperly & Robinson, 1998; Rakoczy, Bergfeld, Schwarz, & Fiske, 2015; Sprung, Perner, & Mitchell, 2007). For instance, there may be two co-referential descriptions for the very same entity, such as "president of the United States" and "Donald Trump". Maxi believes that the president of the United States wants to build a wall, but Maxi does not know about the other co-referential description (Donald Trump). Thus, inferring that Maxi also believes that Donald Trump wants to build a wall would be wrong (fake news, so to say). In such intensionality tasks, a participating child knows, for example, about two descriptions for the very same object that is hidden in a box (e.g., a die is also an eraser). The agent only knows about one of these descriptions (e.g., a die). The agent then sees how the object is transferred to another box under the unknown description (e.g., as an eraser), and children are asked where the agent will look for the object under the known description (e.g., as a die).

Astonishingly, all these different kinds of verbal FB tasks are mastered around the same time and are synchronized and correlated in development (Gopnik & Astington, 1988; Perner & Roessler, 2012; Rakoczy et al., 2015; Wellman et al., 2001). This suggests that children in fact acquire a full-blown, flexible and unified ToM capacity around the age of four. Since these tests use verbal measures to access FB understanding, the underlying capacity is often referred to as *explicit* ToM.

Mindreading before age four

Before children engage in FB representation at around four years of age, they gain an astonishing variety of socio-cognitive skills and the concept of less complex mental states, such as of intention and attention of other people, from the end of the first year on (see e.g., Tomasello, 1999; Tomasello & Rakoczy, 2003). This incidence is often referred to as the "nine-month revolution". Shortly after about nine months of age, infants develop various joint attentional behaviors in synchronized and correlated fashion that necessitate coordinating their actions with objects and other persons, so called triadic interactions (for a comprehensive longitudinal study on various measures, see Carpenter, Nagell, Tomasello, & Butterworth, 1998). For instance, infants follow into adults' attention by following their gaze or gestures, and by

imitatively copying their behavior. At the same time, infants try to get adults into their attention by using deictic gestures, such as showing objects or pointing at objects. As a result of these new behaviors, the infant and the adult share attention to an object for a certain period of time. On the cognitive level, at about the same age, infants understand others' actions as goal-directed and intentional. In a habituation study by Gergely et al. (1995), for example, 12-month-olds that previously watched a ball jumping over an obstacle, looked longer at a novel irrational event (the ball jumped in the absence of an obstacle) compared to a novel rational event (the ball moved straight to the other side). In another interactive study, children selectively imitated the action of a demonstrator that turned on a light switch with his head instead of his hand, when the hands were free, but not when the hands were blocked (Gergely, Bekkering, & Király, 2002). This suggests that infants understood the demonstrator's action as his goal in the first, but as means to a goal in the latter case. In addition, infants at one year of age show an understanding of others' perspective by following gaze around barriers (e.g., Moll & Tomasello, 2004), but also understand which entities others are paying attention to. In a study by Tomasello and Haberl (2003), an infant played with two adults and two toys. During the absence of one of the adults, the other adult introduced a new toy. When the second adult returned, he expressed excitement about the three object aligned on a table and asked the infant, "Can you give it to me?" Infants offered the new toy, suggesting that they understood that people attend to new things, and also that they were able to identify which object was new to the other person.

These early socio-cognitive capacities may be precursor to the later developing FB understanding. Accordingly, in course of the first to fourth year of life, children gradually develop more sophisticated mental state concepts, such as the understanding of simple desires, or of the knowledge-ignorance distinction (that someone does or does not know something), which may even follow a fixed order in development (Wellman & Liu, 2004).

Developing an explicit Theory of Mind

A lot of research has been devoted to developmental determinants and cognitive underpinnings of explicit ToM. Four main predictors have been highlighted in several cross-sectional, longitudinal and intervention studies. First, explicit ToM builds on executive function, in particular working memory (Davis & Pratt, 1995) and inhibition skills (Carlson & Moses, 2001; Rakoczy, 2010; Sabbagh, Moses, & Shiverick, 2006), which may provide the necessary tools to handle and suppress different (and diverging) perspectives of oneself and others. Second,

linguistic development is important for the acquisition of explicit ToM. A meta-analysis on the impact of language on ToM revealed different and complementary developmental roles for general linguistic capacity, as well as semantic, syntactic and pragmatic experience (Milligan, Astington, & Dack, 2007). Studies suggest that engaging in perspective-shifting discourse and using mental state language (such as mental verbs like know, think and believe) may be the crucial linguistic aspects that promote the cognitive construction of mental state concepts (e.g., Lohmann & Tomasello, 2003; Ruffman, Slade, & Crowe, 2002). Evidence that language in fact promotes the development of FB understanding comes from studies on deaf children. Deaf children of hearing parents that are typically delayed in linguistic development due to later exposure to language, perform worse in FB tasks compared to deaf children from deaf families or hearing children (Peterson & Siegal, 1999; Schick, de Villiers, de Villiers, & Hoffmeister, 2007). Additionally, a recent neurophysiological study using functional magnetic resonance imaging (fMRI) found that deaf children with delayed language access showed reduced selectivity in brain regions associated with ToM (Richardson et al., 2018). Third, several findings suggest that socio-interactive experiences, such as attachment security, parental mind-mindedness (caregivers' thinking of their children as individuals with a mind) or sibling interaction, promote ToM development (e.g., McAlister & Peterson, 2007; Meins et al., 2002; Perner, Ruffman, & Leekam, 1994). Fourth, explicit ToM capacities have been found to build on simpler socio-cognitive capacities such as joint attention and action understanding (Aschersleben, Hofer, & Jovanovic, 2008; Licata, Kristen, & Sodian, 2016; Sodian & Kristen-Antonow, 2015; Wellman, Lopez-Duran, LaBounty, & Hamilton, 2008; Wellman, Phillips, Dunphy-Lelii, & LaLonde, 2004).

Different theoretical accounts on the development of an explicit ToM have been established. Theory-theory (Gopnik & Wellman, 1994) suggests that children's changes in the understanding of mind equal theory changes. By observing the behavior of others, children make inferences and form naïve theories about that behavior, which are constantly revised by more sophisticated theories during development. Simulation accounts (Meltzoff & Gopnik, 1993) suggests that children make use of their own mental world as a reference to simulate others' thoughts and feelings, i.e. they understand that others are "like me". And different social interaction accounts (e.g., Carpendale & Lewis, 2004; Liskowski, 2018; Tomasello & Rakoczy, 2003), on the other hand, emphasize the important role of interactive experiences and language for the development of ToM. According to these accounts, infants' special adaptations in joint attentional behavior and the understanding of others' attention and intention in the first year of

life enable them to engage in triadic interactions, which provide the necessary circumstances for further socio-cognitive development.

New findings with implicit measures

Only recently, since about a decade ago, studies using implicit, non-verbal measures of FB understanding revolutionized ToM research. These studies suggest that infants grasp a concept of belief much earlier than previously assumed (see for a review Scott & Baillargeon, 2017). In order to distinguish these findings from those of explicit tasks, the capacity will be further referred to as *implicit* ToM.

Implicit false belief tasks

In 2005, Onishi and Baillargeon published their influential study investigating whether 15-month-old infants show sensitivity to the belief of others by using the violation-of-expectation paradigm. In a scenario that was conceptually based on the classical change-of-location task, infants saw an actress reaching for a watermelon into one of two boxes. In the test conditions, the watermelon changed from one box to the other in self-propelled manner, which was either seen by the actress (true belief (TB) condition) or she was unable to see this event because her view was blocked (FB condition; there were actually even more conditions). Afterwards, the actress reached into a box and rested with her hand in the box that was either congruent with her belief about the location of the watermelon (the new location in the TB condition, but the former location in the FB condition) or incongruent with her belief (the former location in the TB condition, but the new object location in the FB condition). Infants in both test conditions looked significantly longer at the incongruent event compared to the congruent event. This looking pattern suggests that infants formed an expectation on the behavior of the actress based on her mental (mis-)representation of the situation. The findings by Onishi and Baillargeon, thus, challenge the classical view that FB representation develops after the fourth birthday and re-date the age of emergence into the second year of life.

Inspired by these new findings, a variety of non-verbal FB tasks with different measures were established. For instance, anticipatory looking tasks measure whether infants reveal in their looking behavior that they anticipate an agent's mistaken action when this agent holds a FB about an object location (Clements & Perner, 1994; Southgate, Senju, & Csibra, 2007; Surian &

Geraci, 2012). In the study by Southgate et al. (2007), infants watched a movie in which an agent sat behind a screen containing two windows, watching a teddy bear moving a toy between two containers, one standing in front of each window. The teddy put the toy into one container (first object location), which was witnessed by the agent. Then, in one condition, the agent saw how the teddy moved the toy to the other container (second object location; FB1 condition), and in another condition, she was distracted by a phone call during that event (FB2 condition). In both conditions, the agent did not witness how the teddy removed the ball from the second box and disappeared with the toy from the scene (agent distracted by a phone call in both conditions). Thus, in both cases, the agent held a FB about the location of the object (second object location in FB1, first object location in FB2), which was actually not present anymore. Afterwards, the agent turned back at the scene and both windows were illuminated, which indicated that she was about to reach through one of the windows for the toy. 25-month-olds in that study correctly anticipated with their first look at a window (and the duration they looked at each window) in the FB1 condition that the agent would reach for the second object location, and in the FB2, that she would reach for the first object location, which was in each case congruent with the agent's belief. Other studies using this paradigm found FB understanding at even younger ages, at 17 or 18 months of age (Senju, Southgate, Snape, Leonard, & Csibra, 2011; Surian & Geraci, 2012).

Several studies were published that used interaction-based measures of implicit ToM (D. Buttelmann, Carpenter, & Tomasello, 2009; Knudsen & Liszkowski, 2012a, 2012b; Southgate, Chevallier, & Csibra, 2010). One study, for example, came up with the so called "sefo task" (Southgate et al., 2010). In that task, an experimenter showed two novel objects to infants (17-month-olds in this case) and put each object in a separate box. When she shortly left the room in the FB condition, another person entered and interchanged the objects from one box to the other. On her return, the experimenter pointed at one of the boxes and asked the infant to give her the object by using a novel label, sefo. In an adapted TB condition, the experimenter witnessed how the other person swapped the objects. Congruent with the experimenter's belief about the location of the sefo, infants retrieved the object from the non-referred box in the FB condition, but the object from the referred box in the TB condition.

Additionally, other studies reported even more measures of implicit FB understanding in infants, such as neurophysiological signatures of belief-based action prediction using electroencephalography (EEG; Southgate & Verneti, 2014), or emotional face expressions of tension (as e.g., lip biting) during FB stories (Moll, Kane, & McGowan, 2016; Moll, Khalulyan, &

Moffett, 2017). There are even studies to date showing that infants pass non-verbal tasks that measure other FB concepts than change-of-location, such as unexpected-content (D. Buttelmann, Over, Carpenter, & Tomasello, 2014) or intensionality tasks (F. Buttelmann, Suhrke, & Buttelmann, 2015; Scott, Richman, & Baillargeon, 2015), suggesting a rather sophisticated FB competence. A converging line of evidence comes from studies that aim at facilitating performance of children younger than four in standard explicit FB tasks: By reducing linguistic or other task demands, children at age three show enhanced performance or even pass (Mitchell & Lacohee, 1991; Psouni et al., 2018; Rhodes & Brandone, 2014; Rubio-Fernández & Geurts, 2013; Sullivan & Winner, 1993).

Taken together, the past decade provides impressive evidence from around 30 different studies with implicit measures that converge on the claim that even very young infants engage in FB representation (see Scott & Baillargeon, 2017). Interestingly, even great apes, which usually fail FB task (Call & Tomasello, 2008), pass implicit ToM task that are adapted to the infant versions (D. Buttelmann, Buttelmann, Carpenter, Call, & Tomasello, 2017; Krupenye, Kano, Hirata, Call, & Tomasello, 2016).

Theoretical accounts on implicit ToM

Then why is there this vast contrast in the onset of FB understanding between findings of classical explicit task and those of the novel implicit task? Several far reaching theoretical accounts have put forward a controversy regarding the nature of the underlying competencies. Nativist accounts, on the one hand, claim that implicit tasks prove that infants possess a concept of belief similar to that of preschool children (Carruthers, 2013; Leslie, 2005; Scott & Baillargeon, 2017). They suggest that ToM is a domain-specific, modular capacity (mental states are ascribed automatically by a ToM module) and probably even inborn. Infants and toddlers fail standard verbal FB tasks due to performance rather than competence problems, i.e. extraneous task demands, such as linguistic or inhibitory ones, camouflage the ToM competence. Thus, ToM should be operational early in development and independent of experience. Others, on the other hand, doubt that implicit tasks measure proper FB understanding (e.g., Heyes, 2014b; Rakoczy, 2012). Deflationary (or skeptical) accounts suggest that findings of implicit tasks could be interpreted in more parsimonious ways and there might be alternative explanations (Heyes, 2014a; Perner & Ruffman, 2005). Infants could, for example, apply behavior rules or simply react to perceptual novelty. Two-systems accounts assume that humans have two systems to track

mental states of others (Apperly & Butterfill, 2009; Low, Apperly, Butterfill, & Rakoczy, 2016). Explicit ToM tasks tap a flexible, cognitively effortful and full-blown ToM system. Implicit tasks, on the contrary, tap an automatic, efficient and early-developing ToM system that is capable of tracking belief-like states. Two-systems accounts predict that automatic ToM competencies reveal signature limits on the complexity of ToM processing to achieve efficiency. Regarding signature limits, it has been suggested that only a full-blown ToM would be capable of representing the intentionality of others' beliefs. This view is supported by findings showing superior performance of infants in implicit change-of-location tasks as opposed to implicit intentionality tasks (Fizke, Butterfill, van de Loo, Reindl, & Rakoczy, 2017; Oktay-Gür, Schulz, & Rakoczy, 2018). Further evidence for two mindreading systems can be found in studies showing that even adults engage in automatic, unconscious belief-tracking (Kovács, Téglás, & Endress, 2010; Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010; Schneider, Bayliss, Becker, & Dux, 2012). A recent shared intentionality account (Tomasello, 2018) suggests that infants (and apes) track what another person has seen and knows, to predict the other's behavior and pass implicit tasks. They do not take into account their own perspective or the objective situation, and thus, show no FB representation. Only when they come to understand that subjective representations can differ from an objective situation (via experiencing triadic social interactions; at about age four), they have a true concept of belief.

Rationale of the current thesis

The objective of the current thesis was to systematically investigate implicit ToM capacities: Testing whether implicit ToM is a real and robust phenomenon and arriving at a comprehensive characterization of the limits and developmental origins of these early capacities. According to the presented background, three main questions are guiding this investigation.

Research questions

The first question concerns the reliability of findings from implicit ToM tasks. The various findings of these tasks have been taken as evidence for far reaching theoretical accounts, which all converge on the basic assumption that existing findings are reliable. Yet, we do not know how robust and replicable the findings really are. Questions about the reliability of effects have become more and more pressing in psychological science during the last years, because

systematic replication studies revealed that classical findings could often not be reliably reproduced (Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013; Makel, Plucker, & Hegarty, 2012; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons, 2014). Studies investigating an explicit ToM capacity around age four have provided evidence of strong reliability and robustness (e.g., Wellman et al., 2001). However, hardly anything was known about the replicability of existing implicit ToM findings. Based on recent studies in the past year, we now know that most measures could not be replicated in independent labs (see e.g., Crivello & Poulin-Dubois, 2018; Kulke & Rakoczy, 2018; Powell, Hobbs, Bardis, Carey, & Saxe, 2018; Schuwerk, Priewasser, Sodian, & Perner, 2018). This is an unfavorable situation, since most of the original studies used rather small sample sizes and single trial designs, making them vulnerable to spurious or false-positive findings. Additionally, there is currently no meta-analysis on implicit ToM and we do not know about a potential body of unpublished replication studies. In fact, a recent survey made a start and revealed a lot of so called “file drawers”, data that did not make it into publication, suggesting a variety of partial and failed replication attempts on implicit ToM tasks (Kulke & Rakoczy, 2018).

The second question is whether performance on implicit ToM tasks reflect a unitary cognitive capacity. Studies using explicit measures of FB understanding have shown unity for explicit ToM, i.e. several superficially different tasks showed convergence and correlation (Gopnik & Astington, 1988; Perner & Roessler, 2012; Rakoczy et al., 2015). In the past decade, research has accumulated evidence for infants’ and toddlers’ FB understanding from various implicit measures (see Scott & Baillargeon, 2017). They are surprised when an agent acts contrary to his FB in violation-of-expectation tasks (Onishi & Baillargeon, 2005; Träuble, Marinović, & Pauen, 2010), correctly anticipate an agent’s action who is mistaken about an object location in anticipatory looking tasks (Clements & Perner, 1994; Southgate et al., 2007), and offer appropriate helping behavior for agents in interactive FB tasks (D. Buttelmann et al., 2009; Southgate et al., 2010). Each of these different tasks and measures itself may reflect a conceptual ToM capacity, however, for each local finding there may also be alternative “low-level” explanations (e.g., Heyes, 2014a). Accordingly, if the different tasks indeed all tap the same underlying competence, implicit ToM as (to some degree) two-systems accounts would suggest, or proper ToM as nativists would suggest, there should be convergence and correlation between them. Skeptical accounts, on the other hand, would not predict unity between implicit measures. Regarding convergent validity of the various implicit measures, there are hardly any systematic

studies. One longitudinal study found a correlation between an implicit anticipatory looking task and a later explicit ToM task (Thoermer, Sodian, Vuori, Perst, & Kristen, 2012). Another study found no evidence for convergent validity between a violation-of-expectation measure and a helping measure of implicit ToM (Poulin-Dubois & Yott, 2018). Thus, systematic studies on the convergent validity of implicit ToM measures are highly required.

A third question concerns with the developmental origins and cognitive underpinnings of implicit ToM competencies. A large body of studies has revealed comprehensive insight into developmental determinants of explicit ToM: Explicit ToM builds on executive function (e.g., Rakoczy, 2010), different forms of language (e.g., Lohmann & Tomasello, 2003), socio-interactive experience (e.g., Meins et al., 2002) and simpler socio-cognitive capacities (e.g., Wellman et al., 2004). For implicit ToM, however, no longitudinal but only few cross-sectional studies exist, which revealed a mixed pattern of findings concerning a potential role of executive functions (Grosse Wiesmann, Friederici, Singer, & Steinbeis, 2017; Yott & Poulin-Dubois, 2012) or language (e.g., Low, 2010; Meristo et al., 2012). Thus, empirical data about foundations and determinants of the development of implicit ToM are still outstanding.

Work program

The first question, regarding robustness and replicability of individual implicit ToM tasks, will be addressed in basically all four studies presented in this thesis. In study 1, we conducted conceptual and direct replications of three of the main paradigms of implicit FB understanding in infants: Violation-of-expectation (in a new task conceptually based on Onishi & Baillargeon, 2005), anticipatory looking (with stimuli from Southgate et al., 2007) and interactive helping (Southgate et al., 2010). In studies 2 to 4, we focused on the replicability of different interactive FB measures. Study 2 dealt with the “sefo task” (Southgate et al., 2010), where we conducted direct replications at different age groups, as well as pragmatically modified task versions, to achieve a clear picture of the robustness of the original findings. In study 3, we replicated a narrative ToM task held to reveal FB understanding at age 3 (the Duplo task, Rubio-Fernández & Geurts, 2013). We compared this task to matching control conditions to overcome limitations of the original study. And finally, in study 4, we administered an anticipatory correcting paradigm, which may be amenable to simpler explanations than FB understanding, such as the knowledge-ignorance distinction, and appears to be more promising regarding replicability (Knudsen & Liszkowski, 2012b; Powell et al., 2018).

Regarding convergent validity of implicit ToM tasks, as taken up in the second question, study 1 investigated whether there is evidence for a coherent and unitary implicit ToM competence. While skeptical accounts would predict no unity, and hence unrelated performances in different FB tasks, nativists and two-systems accounts, would predict full unity and correlated performances across FB change-of-location tasks. To investigate whether findings on implicit ToM reflect a unified social-cognitive capacity, just like later-emerging explicit ToM capacities, we conducted several implicit FB tasks (violation-of-expectation, anticipatory looking, interactive helping) in a within-subjects design and tested for correlated performances across the different measures. Additionally, in studies 2 and 3, we compared performance in implicit tasks to that in standard explicit tasks to validate whether there is in fact a distinction between the two ToM capacities, or whether implicit tasks actually measure the same capacity as explicit tasks. In study 3, we additionally correlated performance in a narrative change-of-location task with performance in a task measuring another FB concept. That is, we designed a new intentionality version of the narrative task to access whether there are signature limits of the early ToM competence, as two-systems accounts would suggest (e.g., Low et al., 2016), or whether the ToM competence is as unified and sophisticated as that of older children (e.g., Rakoczy et al., 2015).

Study 4 addresses the third question, whether there are developmental determinants of implicit ToM. We used existing longitudinal data collected between 8 and 14 months of age (by researchers in our lab for another study) and tested the same infants again at two years of age using an anticipatory correcting paradigm to tap the implicit ToM capacity (Knudsen & Liszkowski, 2012b). The longitudinal study provided us with predictor variables of infants' socio-cognitive capacities, as well as of socio-interactive experience accessed in the lab and in natural situations during home visits. While nativists would suggest that early ToM capacities are independent of experience, social interaction accounts (e.g., Carpendale & Lewis, 2004; Liszkowski, 2018) would suggest that socio-interactive experience promotes the understanding of others' minds. A developmental link between early interactive experience and FB understanding has been amply documented for explicit ToM tasks (e.g., Meins et al., 2002; Sodian & Kristen-Antonow, 2015), but no one has looked at such correlations for implicit ToM, yet.

Study 1: How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures

This study was published in the Journal *Cognitive Development* (Dörrenberg, Rakoczy, & Liszkowski, 2018).

Abstract

A growing body of infant studies with various implicit, non-verbal measures has suggested that Theory of Mind (ToM) may emerge much earlier than previously assumed. While explicit verbal ToM findings are highly replicable and show convergent validity, systematic replication studies of infant ToM, as well as convergent validations of these measures, are still missing. Here, we report a systematic study of the replicability and convergent validity of implicit ToM tasks using four different measures with 24-month-olds (N=66): Anticipatory looking, looking times and pupil dilation in violation-of-expectation paradigms, and spontaneous communicative interaction. Results of anticipatory looking and interaction-based tasks did not replicate previous findings, suggesting that these tasks do not reliably measure ToM. Looking time and new pupil dilation measures revealed sensitivity to belief-incongruent outcomes which interacted with the presentation order of outcomes, indicating limited evidence for implicit ToM processes under certain conditions. There were no systematic correlations of false belief processing between the tasks, thus failing to provide convergent validity. The present results suggest that the robustness and validity of existing implicit ToM tasks needs to be treated with more caution than previously practiced, and that not all non-verbal tasks and measures are equally suited to tap into implicit ToM processing.

Introduction

How does our capacity to understand each other as rational agents with an inner life and subjective perspectives on the world, also known as “Theory of Mind” (ToM), develop? An enormous research program in developmental psychology has been devoted to this question over the last decades (Wellman, 2014). Recently, this research has been revolutionized by new studies with novel methods and surprising findings. In contrast to most tasks traditionally used in ToM research that relied heavily on verbal questions, these new studies have developed

completely non-verbal and otherwise simplified, implicit tasks suited for testing even very young infants. The findings from these studies have been received as ground-breaking: They suggest that ToM, in particular the capacity to ascribe false beliefs (FB) to other agents – the litmus test for understanding subjectivity (Wimmer & Perner, 1983) – emerges much earlier than previously assumed in the first months of life (Baillargeon, Scott, & Bian, 2016; Baillargeon, Scott, & He, 2010; Scott & Baillargeon, 2017). A converging line of research suggests that these precocious ToM capacities may remain intact and largely automatic over the lifespan, as indicated by findings that adults often seem to engage in spontaneous yet utterly unconscious ToM processing (Kovács et al., 2010; Samson et al., 2010; Schneider et al., 2012; van der Wel, Sebanz, & Knoblich, 2014). Various kinds of such implicit measures have been used with infants, including looking time used with infants as an indicator of violations of expectation (Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007; Träuble et al., 2010), anticipatory looking (Clements & Perner, 1994; Southgate et al., 2007; Surian & Geraci, 2012) and interactive measures such as spontaneous helping (D. Buttelmann et al., 2009; Knudsen & Liskowski, 2012a, 2012b; Southgate et al., 2010). These studies have produced evidence that in their spontaneous looking and interaction behavior, even very young infants seem capable of engaging in FB representation.

From a theoretical point of view, these findings have been taken as evidence for far-reaching theoretical accounts. According to nativist accounts, the findings suggest that ToM is a domain-specific, probably modular, capacity which is online very early in ontogeny and probably even inborn (e.g., Carruthers, 2013; Leslie, 2005). Standard verbal tasks have failed to uncover these early ToM competencies due to extraneous (linguistic and/or inhibitory) performance factors of the tests. According to recent two-systems accounts, the positive findings from the new implicit tasks reflect an early-developing, evolutionarily more ancient, largely automatic and efficient mindreading system. This system is distinct from and potentially the developmental basis for the later-developing, fully-fledged explicit and flexible ToM system tapped in classical verbal tasks (Apperly & Butterfill, 2009; Low et al., 2016).

Robustness, reliability and replicability of implicit ToM tasks

From an empirical point of view, however, it is still unclear how robust, reliable and replicable these results from the novel implicit measures really are. Questions of reliability and replicability of experimental findings have recently taken center-stage in methodological debates

about the evidential status of psychological research (Bakker et al., 2012; Button et al., 2013; Makel et al., 2012; Simmons et al., 2011; Simonsohn et al., 2014). In this context, systematic replication attempts across many labs often yield negative results such that existing, often classical, effects cannot be robustly reproduced in independent labs (Open Science Collaboration, 2015). As a consequence, the value and necessity of large-scale and systematic replication studies are now virtually ubiquitously acknowledged in cognitive psychology. In research on automatic ToM in adults, questions of replicability and interpretation of existing results with implicit tasks have recently begun to be addressed (Heyes, 2014b; Kovács, Téglás, & Endress, 2016; Phillips et al., 2015; Schneider, Slaughter, & Dux, 2017).

Surprisingly, however, hardly anything is known to date about the robustness, reliability and replicability of implicit ToM findings in infants. This is surprising since reliability issues may be particularly pressing in this area of research: First of all, there are still relatively few established infant studies from implicit measures with positive findings, and most of the published studies have used rather small sample sizes and single trial designs, making them vulnerable to spurious findings (Scott & Baillargeon, 2017). Second, to date there are no meta-analyses and we currently do not know about the potential body of unpublished failed replication attempts (the so-called *file-drawer problem*). Third, for most of the published studies there have not been any published replications in independent labs. Fourth, in the few exceptional cases where there are published replication attempts (though they are mostly conceptual, and not direct replications, and often administer multiple within-subject conditions), results are often negative (Grosse Wiesmann et al., 2017; Poulin-Dubois & Yott, 2018; Thoermer et al., 2012; Yott & Poulin-Dubois, 2016; Zmyj, Prinz, & Daum, 2015).

Convergent validity of implicit ToM tasks

A second fundamental question regarding implicit ToM findings in infants concerns their interpretation and validity. Even if individual implicit ToM tasks turned out to be reliable, this would still not settle issues of validity. What is needed are tests of the convergent validity of individual paradigms. If different tasks are in fact all tapping the same underlying cognitive phenomenon – implicit ToM – then they should converge and correlate. Such correlational patterns of superficially different tasks all designed to tap the same underlying phenomenon have been amply documented for explicit ToM (Astington & Gopnik, 1988; Hamilton, Brindley, & Frith, 2009; Perner & Roessler, 2012; Rakoczy et al., 2015). For implicit ToM, however, there

hardly have been any analogous studies of convergent validation by correlation. One recent study has investigated diachronic correlations between infant implicit and later explicit ToM measures in a longitudinal design (Thoermer et al., 2012). In this study, an implicit measure (anticipatory looking) in a very specific type of FB task (change-of-location) predicted performance in later explicit FB tasks, but only in superficially analogous (change-of-location) ones and not in other FB tasks. Given the very local nature of this correlation, however, this finding leaves open different interpretations in rich (implicit tasks tap the same kind of ToM processes as later explicit ones) or lean terms (the shared variance between the tasks is reducible to commonalities in the surface features).

With regard to studies of synchronic correlations of various implicit ToM tasks at a given time, to our knowledge there are so far only two studies from one lab. One study (Yott & Poulin-Dubois, 2016) tested infants in a VoE FB task (conceptually after Onishi & Baillargeon, 2005) and in other implicit tasks of their understanding of desires and intentions. Results revealed that – in addition to not replicating the original FB task finding – there was no systematic pattern of inter-task correlations comparable to those found in explicit ToM tasks. Another study (Poulin-Dubois & Yott, 2018) examined 18-month-olds' performances between different ToM constructs. These included a VoE FB task (conceptually after Onishi & Baillargeon, 2005) and an interactive FB task (conceptually after D. Buttelmann et al., 2009), which both could not be replicated and failed to show any correlations. However, given the diverse, yet un-validated and un-replicated, set of further infant FB tasks, more studies are required that use different tasks to test for convergent validity and the robustness of findings.

Rationale of the present study

Against this background, the rationale of the present study was to test for the reliability and validity of implicit FB measures in infants more systematically and comprehensively. First, in order to examine the robustness and replicability of individual measures, we implemented direct and conceptual replications of structurally very similar implicit ToM tasks, using three different kinds of dependent measures: Anticipatory looking (with the stimuli from Southgate et al., 2007), communicative interaction (Southgate et al., 2010), looking time in a new eye-tracking-based VoE task (conceptually after other VoE studies, Onishi & Baillargeon, 2005; Surian et al., 2007; Träuble et al., 2010). We tested 24-month-olds because (i) this is the youngest age group to perform proficiently in the Southgate et al. (2007) anticipatory looking task, (ii) 20 to 31-month-

olds have been shown to succeed in different VoE tasks (He, Bolz, & Baillargeon, 2011; Scott, 2017; Scott, He, Baillargeon, & Cummins, 2012), and (iii) children have been successfully tested in interactive FB tasks at 2 and 3-years of age (Király, Oláh, Kovács, & Csibra, 2016; Knudsen & Liskowski, 2012b; Rhodes & Brandone, 2014).

Second, we aimed to test for validation of these measures. In a first step, in exploratory ways, we reasoned that if these different tasks all tap the same underlying phenomenon (implicit ToM), then this phenomenon should be measurable in novel ways as well. Much like, for example, infant categorization processing can be tapped in analogous ways by various behavioral and physiological measures (e.g., Elsner, Pauen, & Jeschonek, 2006), infant implicit ToM should reveal itself in various novel behavioral and physiological parameters. In a first step in this direction, a recent study found a novel neurophysiological signature of belief-based action prediction, i.e. mu-desynchronization measured with EEG revealed that infants predicted an agent, who wanted an object from a box but held a false belief about the content of the box, to reach into the empty box, or not to reach into the full box (Southgate & Verneti, 2014). Here, we took a complementary approach by using a novel pupillometrical measure, in addition to looking times, in a VoE FB task. Pupil dilation measures have recently begun to be used in developmental research as another window into the infant mind (Hepach & Westermann, 2016). Increase in pupil size, if not due to luminance, typically indicates arousal and heightened levels of attention. We thus reasoned that violations of expectation should lead to heightened levels of attention and an increase in pupil size, which should correlate with the pattern of looking times (Gredebäck & Melinder, 2011; Jackson & Sirois, 2009). Following this logic, we created a new eye-tracking-based VoE FB task, where we showed infants scenarios in which an agent had a false belief and then acted either in belief-congruent or belief-incongruent ways (looking for the object where he falsely believed it to be or where it really was, respectively) and measured both looking times and pupil dilation. If infants really engage in spontaneous implicit FB processing and thus expect the agent to search in belief-based ways, they should look longer, and show increased pupil dilation in response to the unexpected outcome. Measuring looking times with an eye-tracker has been established in various VoE studies (e.g., Jackson & Sirois, 2009; Köster, Ohmer, Nguyen, & Kärtner, 2016; Yeung, Denison, & Johnson, 2016). We also measured pupil size changes in response to the induction of true beliefs (TB) versus false beliefs, because witnessing the induction of a false belief might already lead to heightened arousal, as has been shown in the emotional expressions of slightly older children (Moll et al., 2016, 2017).

In a second step to test the validity of different implicit FB tasks, we examined their convergent validity by testing for correlations between the four different measures more generally: Anticipatory looking (Southgate et al., 2007), looking times as an indicator of violation-of-expectation as well as pupil dilation (in our new eye-tracking-based VoE task), and spontaneous communicative interaction (Southgate et al., 2010). Since these are all implicit ToM tasks that were specifically designed to reduce processing load and were mastered by the majority of infants in previous studies, we assumed that differences in task demands across tasks should be minimal. Thus, if these tasks indeed all conceptually measure the same, namely implicit forms of representing an agent's belief, as assumed by early mindreading accounts and to some extent by two-systems accounts, then they should converge and correlate.

To replicate previous findings, we were careful to include single trial analyses in all our tasks and measures, as in the original studies. Therefore, we made sure our within-subject sample was sufficiently large to allow for between-subject analyses of the first trial of each task. In addition, we ran within-subject analyses across several trials, because these analyses are based on the larger sample and have more power. Further, we made sure to test for predicted effects directly with planned comparisons, and report one-tailed results for the planned comparisons when appropriate. While this is a more lenient procedure, it would make it more likely to replicate previous findings. To validate previous findings, we looked for correlations across the different measures, and in a more exploratory step, at selective measures and composite scores that were most relevant given the pattern of findings.

Material and methods

Participants

66 German 24-month-olds (median age = 24 months; 16 days; age range = 24;4 – 25;0; 36 girls and 30 boys) from mixed, mostly middle-class, socioeconomic backgrounds in the metropolitan city Hamburg were recruited from a databank of children whose parents had previously agreed to participate in infant studies. All infants participated in the Anticipatory Looking (AL) task; 35 of the infants were tested in the false belief conditions of the Anticipation + Outcome (A + O) task and the Interaction task; and 31 of the infants were tested in the true belief conditions of the A + O task and the Interaction task.

When gaze samples in the video-based tasks were below 70%, we rated the quality based on the gaze replay to reduce the drop-out rate. We included those participants only if we had gaze data during all relevant events (e.g., Teddy changing ball locations, agent reaching through door in outcome phase). After these data reduction steps, our participants had mean weighted gaze samples of 85% (SD = 14) in the AL task and of 80% (SD = 17) in the A + O task. To rule out effects of the amount of tracked gaze samples, we report additional analyses where we only included participants with higher amounts of gaze samples (see Table 2, Appendix 2).

Nine infants were excluded from the AL task because of poor gaze data quality (6) or because they showed no anticipatory looks (3), which resulted in an N of 57 infants. In the A + O task, nine infants were excluded from the false belief condition and five from the true belief condition because of poor gaze data quality (11), fussiness (2) or experimenter error (1), which resulted in an N of 26 infants per condition. In the Interaction task, three infants were excluded from the false belief condition (resulting in an N of 32 infants) and three from the true belief condition (resulting in an N of 28 infants) because they refused to participate.

Design

All infants were tested in three different non-verbal change-of-location paradigms in the following order: a video-based AL task with two false belief versions (original stimuli of Southgate et al., 2007), an interaction-based task with a true belief condition or a false belief condition (adopted from Southgate et al., 2010) and a new eye-tracking-based VoE task that included an anticipation phase comparable to Southgate et al. (2007) but in addition a belief-congruent and a belief-incongruent outcome and a true belief condition or a false belief condition (A + O task). True and false belief conditions were administered between subjects to avoid confusions and longer testing. We chose the least exhausting and biasing task order: first, the AL task was the shortest task and never showed a belief-based outcome, so that it could not reduce belief-based action expectations for the next tasks; second, because calibration ceases over time, we could not administer the AL and A + O tasks back-to-back. The Interaction task was more like a fun game with interesting toys and thus served as a natural break between the eye-tracking tasks; third, the A + O task included belief-incongruent outcomes which could confuse infants or affect their expectations of the agent's actions in later tasks and therefore needed to be administered last. Keeping the task order the same also reduces irrelevant variation which is advisable for correlational analyses.

Set-up and Procedure

Video-based tasks

Eye-tracking set-up

Infants were fastened in a car seat with a headrest to minimize mobility, and watched film clips (25fps, 1280x1024pixel) on a 24inch screen (Dell U2412M) from a viewing distance of approximately 65cm. Display resolution was set at 1920x1200pixel. The screen was surrounded by 2.5m high walls made of black stage cloth. The room had no windows, and room luminance (emitted from the ceiling) was kept constant across all tasks and participants. Sound was played via powered speakers that were hidden behind the screen. Parents were seated centrally behind the infants and were instructed not to interact. Twelve infants sat on their parents' laps because they refused to sit in the car seat. A Tobii (Stockholm, Sweden) X120 eye-tracker was installed underneath the screen and recorded infants' eye movements with a sampling rate of 120Hz. Stimulus presentation and recording were controlled via a Dell Latitude E6530 notebook using Tobii Studio software. We used a 5-point infant calibration. Between the test trials, we showed infants 6s long attention getter videos depicting fun cartoons, e.g. a train or cute bugs, which emitted sounds, to keep attention on the screen.

Anticipatory Looking task

For the Anticipatory Looking task, we used the original video clips of Southgate et al. (2007). For a detailed description of this task see Appendix 1. Infants watched an agent repeatedly reach through one of two windows for a toy hidden in one of two boxes. In the FB1 condition, the agent then witnessed a teddy changing the toy from box 1 to box 2, but did not witness the teddy thereafter removing the toy from the scene (the agent falsely believed the object to be in box 2). In the FB2 condition, the agent did not witness the teddy changing the toy from box 1 to box 2 and then remove it from the scene (the agent falsely believed the object to be in box 1). If infants' anticipatory looking was belief-based, in the FB1 condition infants should anticipate the agent to reach for the location where the object was last ("box 2"); and in the FB2 condition, they should anticipate the agent to reach for the location where the object was first ("box 1"). Since the FB1 condition could simply elicit looks to the ball's last location, FB2 controls for this issue, because the ball's last location is different from the agent's belief of the ball location.

Different to the original study, we mirrored the video clips in order to counterbalance between the target sides. Mirroring had no influence on infants' performances in both test conditions (see Appendix 2 for analyses). Further, before the two familiarization trials, we showed infants two warm-up trials, in which infants saw the agent reaching through the door for a red whale toy that was sitting on a box (one trial on each box), a procedure that has been used by the authors for the same task in follow-up studies (Senju et al., 2010, 2011; Senju, Southgate, White, & Frith, 2009). In contrast to Southgate et al. (2007) who used a between-subject design, all infants in our study saw both the FB1 condition and the FB2 condition in counterbalanced order. A between-subject analysis was still possible by using the first trial performance.

Anticipation + Outcome task

For our new eye-tracking-based VoE task, we recorded video clips based on the stimuli of Southgate et al. (2007), but included a belief-congruent or a belief-incongruent outcome and a FB and a TB version, as in other VoE tasks (Onishi & Baillargeon, 2005; Surian et al., 2007; Träuble et al., 2010). The videos thus contained familiarization trials, a belief-induction phase, and belief-incongruent outcomes. The agent disappeared during the belief-induction, and the desired object remained in the changed location. For a detailed description of this task see Appendix 1. After infants had watched the object placements, the agent duck behind the screen, as in other anticipatory looking studies where the agent disappears before the anticipation phase (Clements & Perner, 1994; Surian & Geraci, 2012; Thoermer et al., 2012), and infants saw an anticipation phase comparable to Southgate et al. (2007) followed by a fixation cross that appeared between the two doors before the outcome to reorient infants' gaze before the outcome phase started. All infants saw a belief-congruent reaching trial and a belief-incongruent reaching trial in counterbalanced order (thus yielding two anticipation trials per infant). We used an eye-tracker to measure looking times, instead of coding live by hand. However, the general procedure was similar: Figure 1 shows selected scenes from the video clips describing the main events. Note, all infants saw the same film clips during the outcome phase (target side counterbalanced left or right) to ensure that there were no spatial or luminance differences between outcomes or conditions that could affect pupil size.

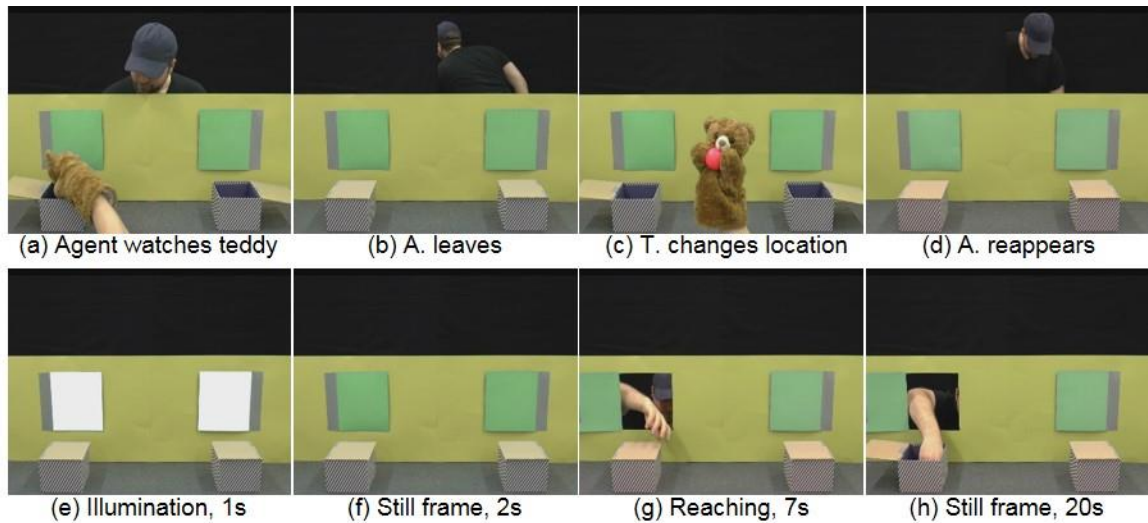


Figure 1. Selected scenes showing the main events in a false belief trial of the Anticipation + Outcome task in consecutive order (a) – (h). (a) Agent watches teddy during the first change of ball location; (c) Teddy changes ball location; (d) Agent reappears before (TB condition) or after location change (FB condition; depicted is the FB condition); (e) Agent ducks behind the screen before illumination; (g, h) Outcome phase, either belief-congruent or belief-incongruent.

Interaction task

Set-up

The testing room was 3.7m x 3.5m in size, had white walls, a door in one corner and three cameras in the other corners recording the experimental procedure. Infants were seated on the floor in front of their parents, who leaned against a wall. A blue and a green box (L: 27cm, H: 34cm, W: 20cm) were placed 120cm from the infant and 100cm apart. The front of the boxes facing the infants could be opened so that they remained in an upward position. For the two warm-up trials, we used a yellow bath duck and a small yellow shovel as objects. Figure 2 shows the three different object pairs we used in the three test trials: (1) a purple lemon squeezer and a red funnel, (2) a black watering can spout with colorful glue strips and a yellow plastic toy ring with colorful glue strips, and (3) a purple pastry scraper and a grey piece of tube with colorful glue strips. For each pair we used a different novel label for the target object: (1) Sefo, (2) Toma, and (3) Nari.



Figure 2. Object pairs used in the three test trials of the Interaction task.

Procedure

We adopted the experimental procedure of the Sefo-task in a FB and a TB version (Southgate et al., 2010, Experiment 1). An experimenter requested infants to retrieve one of two objects he either correctly or falsely believed to be in one of two boxes. In the TB condition, the agent witnessed another person swapping the objects, thus infants were required to retrieve the toy from the box that was indicated; in the false belief condition, on the contrary, the agent was outside during the object changes, infants were thus required to retrieve the toy from the opposite box which the experimenter indicated. For a detailed description of this task see Appendix 1. Instead of one test trial, we administered three test trials in the same condition (either TB or FB) to gather continuous data for our correlational analyses. A first trial between-subject analysis was still possible.

Coding and analyses

For the video-based tasks, the two doors served as areas of interest (AOI) for the analyses during the anticipation phase. We measured first fixations starting onset of the illumination until 1.75s after offset of the illumination (2.75s in total) using Tobii Studio software (I-VT fixation filter). Infants could score 1 (first fixation in AOI of correct door; AL task: FB1 = last object location, FB2 = first object location; A + O task: FB = empty box, TB = full box) or 0 (first fixation in AOI of incorrect door; AL task: FB1 = first object location, FB2 = last object location; A + O task: FB = full box, TB = empty box). For the looking times between the two doors in the anticipation phase, we analyzed raw data using customized R scripts, and measured from onset of the illumination until 1.75s after offset of the illumination (2.75s in total), as authors did in follow-

up studies (Senju et al., 2010, 2011, 2009). We calculated differential looking scores (DLS) for the anticipation phase by subtracting the looking time to the incorrect door from the looking time to the correct door and dividing it by the sum of both (a value of 0 would indicate no preference for one door; a value above 0 would indicate longer looking to the correct door; a value below 0 would indicate longer looking to the incorrect door). To compare results to the original analysis of Southgate et al. (2007), we also analyzed the total looking time instead of the DLS, and used the second familiarization trial as an inclusion criterion for the analyses of the test trials. Due to some ambiguity in the original Southgate et al. (2007) description of the analyses and stimuli, we report additional analyses of our measures for the time period of 1.75s starting after offset of the illumination for the AL task in Appendix 2.

In the A + O task, we measured looking times and mean pupil size during the outcome phase (reaching phase plus still frame phase, 27s) and pupil size additionally at a time point at the beginning of the outcome, in the fourth second of the reaching phase. Note that the pattern of looking time results remained the same when analyzing only the still frame phase of the outcome phase (without the reaching phase). We used the total screen (size of video) as AOI and analyzed raw data using customized R scripts. Mean pupil size of left and right eye was computed at each sample. To analyze changes in pupil size, we calculated the relative change in pupil dilation (RCD) by subtracting a baseline from the focal phase and dividing it by the baseline (baselines are described in the result section). Tobii Studio calculates pupil size by taking distance to the stimuli into account (Tobii AB, 2016). To make sure, minor changes in infants' posture did not affect pupil size calculation, we ran subsidiary analyses (see Appendix 2). First, we used the distance scores from Tobii as the dependent variable in the same manner as we did for the pupil dilation, to see if distance alone would yield similar results as pupil size change. Further, we correlated pupil size and distance to the eye-tracker. Both analyses confirmed that pupil size was not influenced by posture changes. Posture changes were minor anyhow, because infants were fastened in a car seat with a headrest.

In the Interaction task, we coded the box that was first approached or pointed to by the infant. Infants could score 1 (referred box; correct in TB, but incorrect in FB) or 0 (non-referred box; incorrect in TB, but correct in FB). We calculated a mean performance over the repeated trials, ranging from 0 (no trial correct) to 1 (all trials correct). A subsample of 12 FB and 12 TB participants (a total of 72 trials) was additionally analyzed by a second coder. Inter-rater reliability was excellent (Cohen's $k = .944$, $p < .001$).

To analyze relations between the measures, we used Pearson correlations for metric variables, and phi correlations for dichotomous variables. All statistical tests were performed in IBM SPSS Statistics Version 23. Alpha was set at .05. All presented p -values are two-tailed if not mentioned otherwise. We report lower and upper limits of 95% confidence intervals (CI).

Results

For each of our four measures we first report the first trial between-subject performance, to compare it to the original studies. To be most lenient in achieving replication results, we report one-tailed analyses for those comparisons that have revealed significant results in previous studies. We then report analyses on the full set of our data, to seek for confirmatory support with the larger sample. Where appropriate we use order as between-subject factor and report results for specific orders. Because previous findings predict effects specifically for false belief conditions, we analyze conditions also separately.

Anticipatory Looking task

Familiarization trials

48 participants provided data for the first familiarization trial, 51 for the second familiarization trial, and 43 infants for both trials. Across the two familiarization trials, the first look was significantly more often directed to the correct door than expected by chance ($M = .605$, $SD = .279$; $t(42) = 2.46$, $p = .018$, $d_z = .38$, $CI: .019, .191$). In the first trial, 58% of the infants directed their first look to the correct door (binomial test, $n = 48$, $p = .312$, odds ratio (OR) = 1.38). In the second trial, infants directed their first look significantly more often to the correct door than expected by chance (65% correct; binomial test, $n = 51$, $p = .049$, OR = 1.86). There was a negative correlation between infants' first look in the first familiarization trial and in the second familiarization trial ($\phi(43) = -.361$, $p = .018$). Analyses with the DLS as dependent measure revealed a similar pattern. Infants tended to look longer to target than distractor across the two trials ($M = .12$, $SD = .44$; $t(42) = 1.84$, $p = .073$, $d_z = .28$, $CI: -.012, .258$); and in the second trial ($M = .19$, $SD = .71$; $t(50) = 1.87$, $p = .068$, $d_z = .26$, $CI: -.014, .387$); but not in the first trial ($M = .07$, $SD = .79$; $t(47) = .64$, $p = .527$, $d_z = .09$, $CI: -.157, .302$). Also for the DLS, the first and the second familiarization trial tended to correlate negatively ($r(43) = -.296$, $p = .054$).

Anticipation phase

First trial analyses

To exclude any effects of repeated trial exposure, we analyzed the first trial separately for each of the two conditions, and compared these to each other. When analyzing the first trial of those infants who had looked at the correct door in the second familiarization trial, 14 participants provided data for the FB1 condition and 16 for the FB2 condition.

The right panel of Figure 3 shows the first look results for the replication analysis of the finding by Southgate et al. (2007). In the FB1 condition, 71% of the infants directed their first look to the correct door, which was not different from chance (binomial test, $n = 14$, $p = .090$, one-tailed, OR = 2.45). Contrary to the original study, in the FB2 condition, infants' first look went significantly more often to the incorrect door than expected by chance (13% correct; binomial test, $n = 16$, $p = .004$, OR = 6.69). Further, contrary to the original study, the two conditions differed significantly from each other (Fisher's exact test, $n = 30$, $p = .002$, $\phi = -.600$). The pattern of results for our analyses with measurements starting offset of the illumination was similar, see Appendix 2.

In the FB1 condition, the DLS did not differ from chance ($M = .316$, $SD = .819$; $t(13) = 1.44$, $p = .088$, one-tailed, $d_z = .40$, CI: $-.157, .789$). In the FB2 condition, the DLS revealed that infants looked significantly longer to the incorrect door than expected by chance ($M = -.450$, $SD = .665$; $t(15) = 2.71$, $p = .016$, $d_z = .70$, CI: $-.805, -.095$). Condition differed significantly from each other ($t(28) = 2.83$, $p = .009$, $d_s = 1.07$, CI: $.211, 1.32$).

To be as close as possible to the original analyses, we additionally analyzed the total looking time to each door for the first trial for those who passed the second familiarization trial. A 2x2 ANOVA with window (correct, incorrect) as within-subject factor and condition (FB1, FB2) as a between-subject factor yielded no significant main effects of window ($F(1, 28) = .03$, $p = .428$, one-tailed, $\eta_p^2 = .001$, CI: $-353, 423$) or condition ($F(1, 28) = .39$, $p = .535$, $\eta_p^2 = .014$, CI: $-329, 175$), but a significant interaction between condition and window ($F(1, 28) = 9.78$, $p = .004$, $\eta_p^2 = .259$). Infants looked significantly longer at the correct door ($M = 960\text{ms}$, $SD = 819$) compared to the incorrect door ($M = 332\text{ms}$, $SD = 339$) in the FB1 condition ($F(1, 28) = 5.15$, $p = .031$, $\eta_p^2 = .155$, CI: $61, 1194$), but significantly longer at the incorrect door ($M = 1002\text{ms}$, $SD = 661$) compared to the correct door ($M = 444\text{ms}$, $SD = 551$) in the FB2 condition ($F(1, 28) = 4.64$, $p = .040$, $\eta_p^2 = .142$, CI: $-1087, -27$).

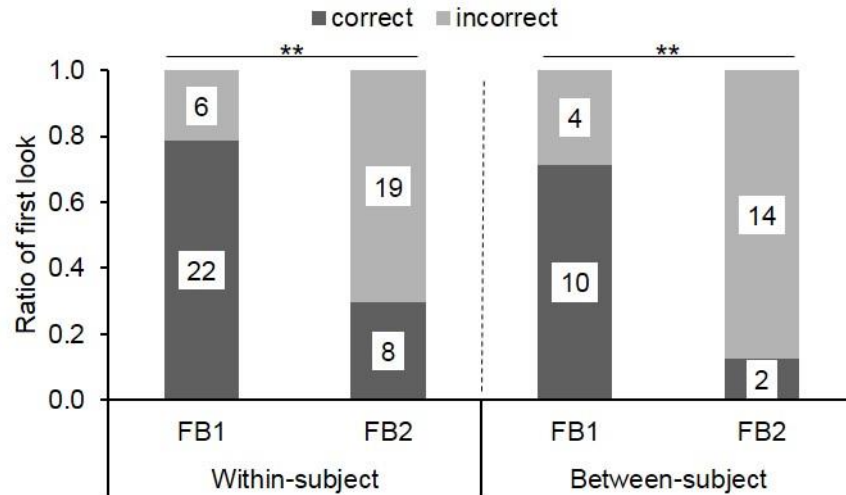


Figure 3. Anticipatory looking task. Percentage of infants who passed the second familiarization trial and first looked to the correct or the incorrect door in the FB1 and the FB2 conditions. The left panel shows a within-subject analysis and the right panel shows the between-subjects replication analysis of Southgate et al. (2007) for the first trial. Numbers in bars show number of infants. $**p < .01$

Analyses of full set

In the test trials, 51 participants provided data for the FB1 condition (50 provided a valid first fixation), and 51 for the FB2 condition (49 provided a valid first fixation). Regarding those infants who passed the second familiarization trial, 29 participants of each condition provided data for the test trials (28 provided a valid first fixation in the FB1, and 27 in the FB2).

When analyzing all infants, in the FB1 condition, 62% directed their first look to the correct door, which is not different from chance (binomial test, $n = 50$, $p = .119$, $OR = 1.63$). In the FB2 condition, infants' first look went significantly more often to the incorrect door than expected by chance (only 24% correct; binomial test, $n = 49$, $p < .001$, $OR = 3.17$). Comparing the two conditions, infants were significantly more often correct in the FB1 condition compared to the FB2 condition (McNemar, $n = 44$, $p = .002$, $OR = 4.82$). Order of condition presentation (FB1 first, FB2 first) had no influence on the FB1 condition (Fisher's exact test, $n = 50$, $p = .383$, $\phi = .155$). In the FB2 condition, performances differed between orders (Fisher's exact test, $n = 49$, $p = .022$, $\phi = -.345$). When FB2 was administered first, infants performed significantly below chance (see first trial analysis). When FB2 was second, infants' first looks were not different from chance (41% correct; binomial test, $p = .523$, $OR = 1.44$).

The left panel of Figure 3 shows the ratio of first looks in each condition for the infants who had correctly anticipated in the second familiarization trial. When analyzing those infants who passed the second familiarization trial, in the FB1 condition, infants' first look went significantly more often to the correct door than expected by chance (79% correct; binomial test, $n = 28, p = .004, OR = 3.76$); in the FB2 condition, infants' first look went almost significantly more often to the incorrect door than expected by chance (30% correct; binomial test, $n = 27, p = .052, OR = 2.33$). Comparing the two conditions, infants were significantly more often correct in the FB1 condition compared to the FB2 condition (McNemar, $n = 24, p = .004, OR = 9.27$).

When analyzing the DLS of all infants, in the FB1 condition, infants looked significantly longer to the correct door than expected by chance ($M = .245, SD = .747; t(50) = 2.35, p = .023, d_z = .34, CI: .035, .455$); in the FB2 condition, infants looked significantly longer to the incorrect door than expected by chance ($M = -.373, SD = .669; t(50) = 3.98, p < .001, d_z = .56, CI: -.561, -.185$). Conditions differed significantly ($t(45) = 3.74, p = .001, d_z = .55, CI: .275, .917$). Order of condition presentation had no influence in the FB1 condition ($t(49) = 1.58, p = .121, d_s = .45, CI: -.741, .089$). In the FB2 condition, performances tended to differ between orders ($t(49) = 1.74, p = .088, d_s = .49, CI: -.049, .690$). When FB2 was administered first, infants performed significantly below chance (see first trial analysis). When FB2 was second, the DLS was not different from chance ($M = -.203, SD = .707; t(23) = 1.41, p = .173, d_z = .28, CI: -.502, .095$).

When analyzing the DLS of those infants who had passed the second familiarization trial, in the FB1 condition, infants looked significantly longer to the correct door than expected by chance ($M = .486, SD = .704; t(28) = 3.72, p = .001, d_z = .70, CI: .219, .754$). In the FB2 condition, infants looked significantly longer to the incorrect door than expected by chance ($M = -.331, SD = .679; t(28) = 2.63, p = .014, d_z = .50, CI: -.590, -.073$). Conditions differed significantly ($t(25) = 3.66, p = .001, d_z = .72, CI: .354, 1.262$).

Anticipation + Outcome task

Looking times in outcome phase

First trial analyses

In a first trial analysis, 26 participants provided data in the FB condition (16 congruent, 10 incongruent) and 26 in the TB condition (14 congruent, 12 incongruent). The right panel of Figure 4 shows the results of the between-subject analysis on the looking time during the outcome phase for the first trial. A univariate ANOVA with condition (FB, TB) and congruency (congruent outcome, incongruent outcome) as between-subject factors provided confirmatory support to previous studies that infants in the group with the incongruent outcome tended to look longer compared to infants in the group with the congruent outcome (main effect of congruency: $F(1, 48) = 2.97, p = .046$, one-tailed, $\eta_p^2 = .058$, CI: -6485, 501), with no significant difference between conditions and no interaction. To test directly the prediction that the effects were present in each condition, simple comparisons based on the variance of the overall ANOVA revealed that infants in the TB condition looked significantly longer when the outcome in the first trial was incongruent compared to when it was congruent ($F(1, 48) = 3.97, p = .026$, one-tailed, $\eta_p^2 = .076$, CI: -9712, 46), but not in the FB condition ($F(1, 48) = .21, p = .323$, one-tailed, $\eta_p^2 = .004$, CI: -6150, 3849).

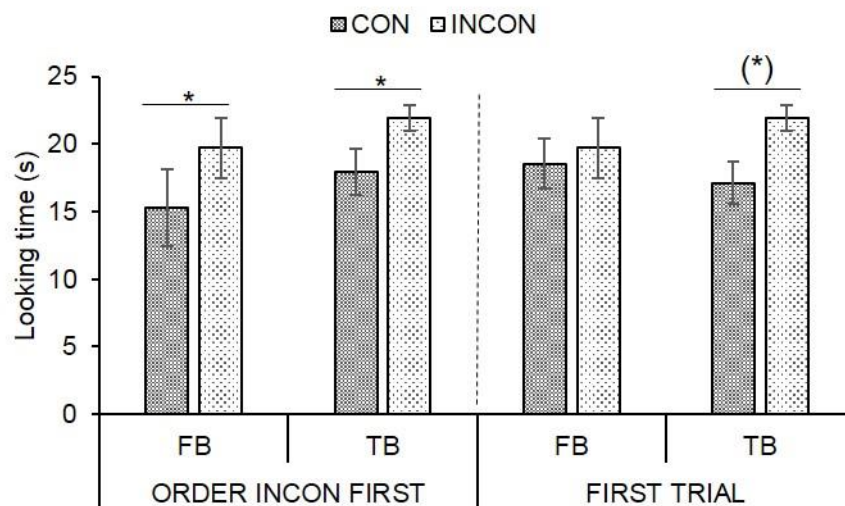


Figure 4. Anticipation + Outcome task. Participants' mean looking time \pm s.e.m. for congruent (CON) and incongruent trials (INCON) in the false belief (FB) and true belief (TB) conditions. Left panel: Within-subject analysis for the order incongruent outcome first ($n = 22$). Right panel: Between-subject analysis for the first trial ($n = 26$ per condition). * $p < .05$; (*) $p = .026$, one-tailed

Analyses of full set

52 infants provided data for both congruent and incongruent trials, 26 per condition. For a within-subject analyses on the outcome phase, a 2x2x2 ANOVA on the mean looking time with congruency as within-subject factor (congruent outcome, incongruent outcome) and condition (FB, TB) and order (congruent first, incongruent first) as between-subject factors revealed a significant main effect of congruency ($F(1, 48) = 5.52, p = .023, \eta_p^2 = .103, CI: -3484, -270$) and an interaction with order ($F(1, 48) = 8.37, p = .006, \eta_p^2 = .149$). Only infants who first saw the incongruent outcome looked significantly longer at the incongruent than the congruent outcome ($F(1, 48) = 11.89, p = .001, \eta_p^2 = .199, CI: -6633, -1746$); infants who first saw the congruent outcome did not look significantly longer at the incongruent than the congruent outcome ($F(1, 48) = 1.76, p = .677, \eta_p^2 = .004, CI: -1653, 2523$). The left panel of Figure 4 shows the mean looking times of infants who saw the incongruent trial first for the congruent and incongruent outcome for both conditions.

To ensure that this effect was present in both conditions, we re-ran the 2x2 ANOVA for each condition. We obtained again significant interactions between congruency and order for both conditions, and simple comparisons based on the overall variance of the ANOVA revealed that in both conditions infants looked longer to the incongruent than congruent outcome when the incongruent trial had been first (FB: $F(1, 24) = 5.22, p = .031, \eta_p^2 = .179, CI: -8387, -427$; TB: $F(1, 24) = 6.95, p = .014, \eta_p^2 = .225, CI: -7084, -864$), but not when the congruent trial had been first (FB: $F(1, 24) = .16, p = .690, \eta_p^2 = .007, CI: -2532, 3761$; TB: $F(1, 24) = .03, p = .856, \eta_p^2 = .001, CI: -2623, 3135$). Figure 5 displays the looking times across the time of the test period for the incongruent first order. The effect appears to emerge most pronounced in the still frame period after 4 seconds and lasts for around 8-10 seconds.

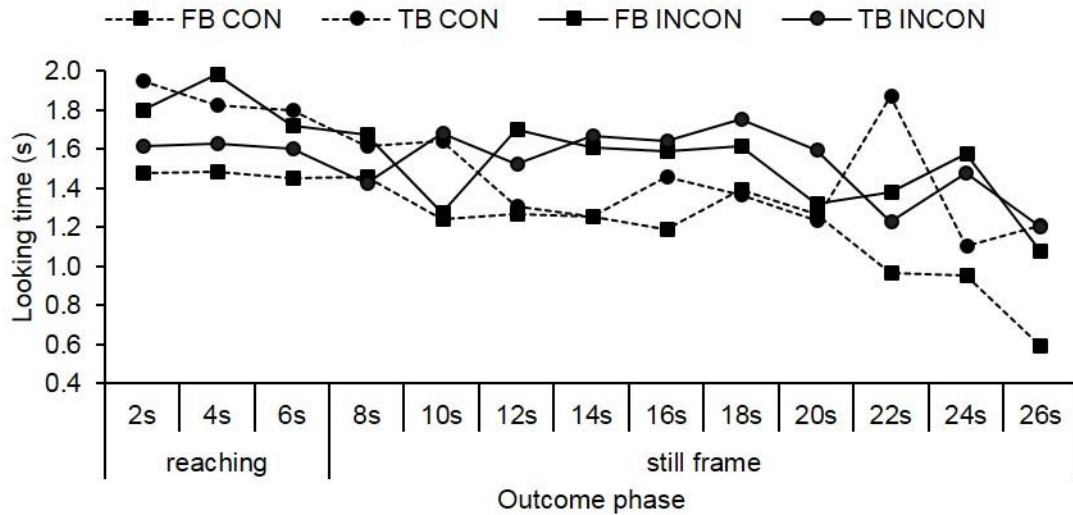


Figure 5. Anticipation + Outcome task. Participants' mean looking time in two-second-increments in congruent and incongruent trials across the outcome phase for each condition in the incongruent outcome first order. First second of reaching phase is not depicted for better illustration.

A similar pattern was obtained when analyzing the number of infants who showed a differential looking pattern across the two trials: 73% of infants who saw the incongruent outcome first looked longer at the incongruent outcome compared to the congruent outcome, which is marginally significant (binomial test, $n = 22$, $p = .052$, $OR = 2.70$). In contrast, only 63% of infants who saw the congruent outcome first looked longer at the incongruent outcome compared to the congruent outcome (binomial test, $n = 30$, $p = .20$, $OR = 1.70$).

Pupil dilation

Outcome phase

We analyzed the RCD from the baseline defined as the mean of the last second of the anticipation phase to the mean of the outcome phase. In a first trial analysis, 25 participants provided data in the FB condition (15 congruent, 10 incongruent) and 26 in the TB condition (14 congruent, 12 incongruent). A between-subject univariate ANOVA with congruency (congruent outcome, incongruent outcome) and condition (FB, TB) as factors revealed that infants in the group with the incongruent outcome had a significantly larger increase in pupil size compared to infants in the group with the congruent outcome (main effect of congruency: $F(1, 47) = 10.54$, $p = .002$, $\eta_p^2 = .183$, $CI: -.083, -.020$), with no significant difference between conditions and no

interaction. To test directly whether the effects were present in each condition, simple comparisons based on the variance of the overall ANOVA confirmed that in the FB condition infants' relative pupil size increase was larger when the outcome in the first trial was incongruent ($M = .112$, $SD = .065$) compared to when it was congruent ($M = .043$, $SD = .052$; $F(1, 47) = 9.10$, $p = .004$, $\eta_p^2 = .162$, $CI: -.115, -.023$); in the TB condition means were in the same direction but did not reach significance (incongruent: $M = .091$, $SD = .069$; congruent: $M = .057$, $SD = .038$; $F(1, 47) = 2.40$, $p = .128$, $\eta_p^2 = .049$, $CI: -.078, .010$).

50 participants provided data for both congruent and incongruent trials, 25 per condition. A 2x2x2 ANOVA for the RCD (from last second of anticipation phase to mean of outcome phase) with congruency (congruent outcome, incongruent outcome) as within-subject factor, and condition (FB, TB) and order (congruent first, incongruent first) as between-subject factors revealed a main effect for congruency ($F(1, 46) = 8.20$, $p = .006$, $\eta_p^2 = .151$, $CI: -.044, -.008$) which interacted with order ($F(1, 46) = 4.87$, $p = .032$, $\eta_p^2 = .096$), with no difference between the conditions. Simple comparisons based on the variance of the overall ANOVA revealed that infants' relative pupil increase was larger in the incongruent ($M = .101$, $SD = .067$) compared to the congruent outcome ($M = .056$, $SD = .057$) when the incongruent trial was presented first ($F(1, 46) = 11.46$, $p = .001$, $\eta_p^2 = .199$, $CI: -.072, -.018$), but not when the congruent trial was presented first (incongruent: $M = .056$, $SD = .062$; congruent: $M = .050$, $SD = .046$; $F(1, 46) = .246$, $p = .622$, $\eta_p^2 = .005$, $CI: -.030, .018$). To ensure this effect was present in both conditions, we analyzed each condition separately. For the FB condition, when the incongruent trial was presented first, the difference in the RCD between the incongruent ($M = .112$, $SD = .065$) and the congruent ($M = .051$, $SD = .062$) outcome remained significant ($F(1, 23) = 7.75$, $p = .011$, $\eta_p^2 = .252$, $CI: -.105, -.016$). Also for the TB condition, when the incongruent trial was presented first, we found a similar pattern (incongruent: $M = .091$, $SD = .069$; congruent: $M = .061$, $SD = .055$; $F(1, 23) = 3.53$, $p = .074$, $\eta_p^2 = .133$, $CI: -.064, .003$). On the level of individual infants the measure did not differentiate between infants who showed a larger RCD in incongruent compared to congruent trials. Figure 6 shows the temporal unfolding of the pupil sizes during the outcome period for congruent and incongruent trials in the TB and FB conditions when the incongruent trial was administered first. Differences seem to arise as early as 3 seconds into the reaching phase and last almost until the end of the testing period.

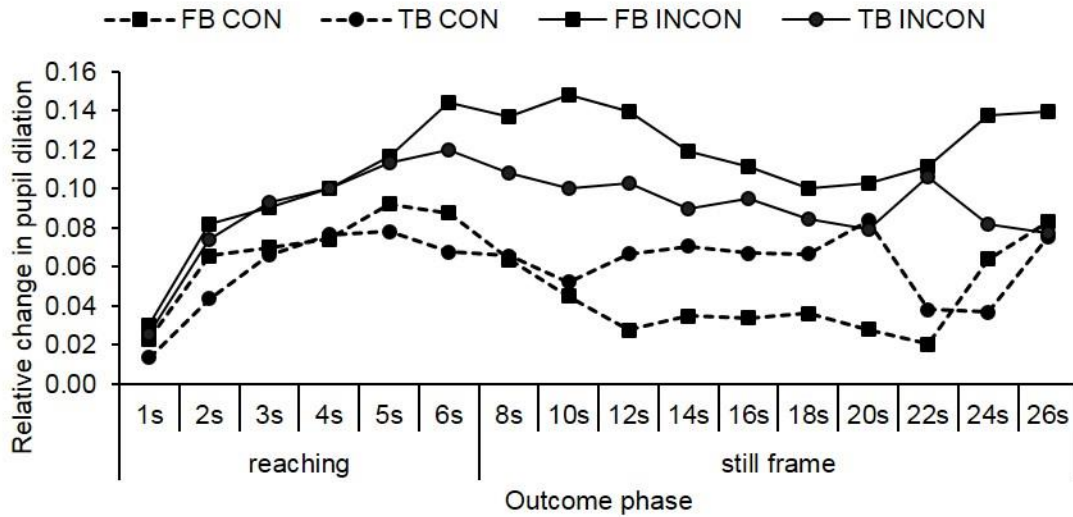


Figure 6. Mean relative change in pupil dilation for participants in the order incongruent outcome first in congruent and incongruent trials across the outcome phase for each condition. First five time points of outcome phase in one-second-increments, all subsequent time points in two-second-increments.

We also tested whether the effect was immediately present following the outcome. If the effect was present early, this would render interpretations that invoke longer looking as one cause of the effect less likely. We focused on the reaching phase when the hand started approaching the container (fourth second of outcome phase, see Figure 6) and re-ran the analyses. The pattern of results with order interactions and selective congruency effects remained the same. Full results are reported in Appendix 2.

Belief induction phase

We also analyzed the mean RCD over the two trials from a pre-induction baseline (the mean of the five seconds before the teddy disappeared and the phone rang) to after the belief induction (the first second after the agent reappeared and sat behind the screen again). 25 participants per condition provided data. A univariate ANOVA with condition (FB, TB) and order (congruent first, incongruent first) as between subject factors revealed a main effect of condition ($F(1, 46) = 18.81, p < .001, \eta_p^2 = .290, CI: .033, .091$) with no effect of order and no interaction. Infants' RCD was significantly larger in the false belief condition ($M = .055, SD = .044$) compared to the TB condition ($M = -.006, SD = .055$). To ensure that there were no learning or other internal effect involved, we analyzed the first trial. The ANOVA yielded again a main effect of condition ($F(1, 41) = 15.59, p < .001, \eta_p^2 = .267, CI: .036, .111$). In the first trial, infants had a larger pupil

size increase in the FB condition ($M = .059$, $SD = .066$) compared to the TB condition ($M = -.015$, $SD = .053$).

To investigate whether this effect was due to the sudden appearance of the agent in the FB condition who had been present in the TB condition already for a longer time, we defined as a new time window for the TB condition the moment when the agent had just appeared and compared it to the moment in the FB condition when the agent had just appeared. A univariate ANOVA with condition and order as between subject factors for the RCD from the pre-induction baseline to the appearance of the agent (early in TB, late in FB) yielded no significant difference between the FB condition ($M = .052$, $SD = .043$) and the TB condition ($M = .045$, $SD = .068$; $F(1, 46) = .20$, $p = .656$, $\eta_p^2 = .004$, $CI: -.025, .039$), suggesting that the effect was based on the appearance of the agent rather than the induction of a false belief.

Anticipatory looking

To provide additional evidence for our analyses in the AL task, we also analyzed infants' anticipatory looking pattern in the current A + O task.

Familiarization trials

51 participants provided data for the first familiarization trial (48 provided a valid first fixation), 47 for the second familiarization trial, and 46 provided data for both trials (44 provided two valid first fixations). Across the two familiarization trials of both conditions, the first look tended to be less often to the correct door than expected by chance ($M = .421$, $SD = .263$; $t(43) = 2.01$, $p = .051$, $d_z = .30$, $CI: -.160, .000$). In the first trial, 52% of the infants directed their first look to the correct door (binomial test, $n = 51$, $p = .885$, $OR = 1.08$). In the second trial, infants' first look was significantly more often directed to the incorrect door than expected by chance (32% correct; binomial test, $n = 47$, $p = .019$, $OR = 2.13$). There was a negative correlation between infants' first look in the first familiarization trial and in the second familiarization trial ($\phi(44) = -.409$, $p = .007$). A similar pattern was obtained with the DLS measure. Infants performed at chance level across the two trials ($M = -.074$, $SD = .552$; $t(45) = .90$, $p = .371$, $d_z = .13$, $CI: -.237, .090$), as well as in the first trial ($M = .033$, $SD = .810$; $t(50) = .29$, $p = .773$, $d_z = .04$, $CI: -.195, .261$). In the second trial, infants looked significantly longer at the incorrect door than expected by chance ($M = -.224$, $SD = .744$; $t(46) = 2.06$, $p = .045$, $d_z = .30$, $CI: -.442, -.005$). There was no correlation between the first and second familiarization trial for the DLS.

Anticipation phase

24 participants provided data in the FB condition, and 25 in the TB condition (24 provided a valid first fixation) for first trial analyses.

In the first test trial, in the FB condition, infants' first looks were not different from chance (63% correct; binomial test, $n = 24$, $p = .307$, $OR = 1.70$); also in the TB condition, infants' first looks were not different from chance (54% correct; binomial test, $n = 24$, $p = .839$, $OR = 1.17$). Conditions were not different from each other (Fisher's exact test, $n = 48$, $p = .385$, $\phi = -.167$). Because there were only few infants who passed the second familiarization trial, we could not use this as an inclusion criterion for further analyses.

The DLS was not different from chance in the FB condition ($M = -.188$, $SD = .602$; $t(23) = 1.53$, $p = .140$, $d_z = .31$, $CI: -.442, .066$), as well as in the TB condition ($M = -.037$, $SD = .625$; $t(24) = .30$, $p = .767$, $d_z = .06$, $CI: -.300, .221$). Conditions were not different from each other ($t(47) = .86$, $p = .395$, $d_s = .12$, $CI: -.504, .202$)

Analyses on repeated trials

In contrast to the AL task which had one trial per condition, in the current A + O task, infants had two test trials per condition. In the second trial, 25 participants provided data in each condition (24 provided a valid first fixation in the TB). 23 participants provided data for both trials in the FB condition, 25 in the TB condition (24 a valid first fixation).

First look. Across the two test trials, first looks were not significantly different from chance in the FB condition ($M = .587$, $SD = .389$; $t(22) = 1.07$, $p = .295$, $d_z = .22$, $CI: -.081, .255$), or in the TB condition ($M = .500$, $SD = .417$; $t(23) = .00$, $p = 1.00$, $d_z = .00$, $CI: -.176, .176$). Conditions were not different from each other ($t(45) = .74$, $p = .464$, $d_s = .22$, $CI: -.150, .324$). This was also true for the second anticipation trial (FB: 56% correct; binomial test, $n = 25$, $p = .690$, $OR = 1.27$; TB: 54% correct; binomial test, $n = 24$, $p = .839$, $OR = 1.17$; Fisher's exact test, $n = 49$, $p = 1.00$, $\phi = -.018$). First looks in the second trial were not different between infants that had first seen a congruent or an incongruent trial (Fisher's exact test, $n = 49$, $p = 1.0$, $\phi = -.036$).

DLS. Across the two test trials, in both conditions the DLS was not significantly different from chance level (FB: $M = -.013$, $SD = .484$; $t(22) = .13$, $p = .896$, $d_z = .03$, $CI: -.223, .196$; TB: $M = .090$, $SD = .454$; $t(24) = .99$, $p = .334$, $d_z = .20$, $CI: -.098, .277$), with no difference between conditions ($t(46) = .76$, $p = .451$, $d_s = .22$, $CI: -.375, .170$). In the second test trial, in the FB condition, the DLS was not different from chance ($M = .223$, $SD = .734$; $t(24) = 1.52$, $p = .141$, $d_z = .31$, $CI: -.080, .526$). However, in the TB condition, in the second test trial, infants tended to

look longer to the correct door ($M = .217$, $SD = .624$; $t(24) = 1.74$, $p = .095$, $d_z = .36$, $CI: -.041, .474$). Conditions were not different from each other ($t(48) = .04$, $p = .972$, $d_s = .01$, $CI: -.381, .394$). There was no difference in the performance of infants in the second trial who had first seen a congruent or an incongruent trial ($t(48) = .79$, $p = .938$, $d_s = .23$, $CI: -.377, .407$). Infants performed significantly better in the second trial compared to the first trial ($t(47) = 2.79$, $p = .008$, $d_z = .40$, $CI: -.601, -.098$).

Interaction task

First trial analyses

To replicate the result by Southgate et al. (2010) we conducted a between-subject analyses and analyzed the first trial. Figure 7 shows the ratio of choice in the first trial for both conditions. In the TB condition, infants chose the referred box significantly more often than expected by chance (71% correct; binomial test, $n = 28$, $p = .036$, $OR = 2.45$). In the FB condition, infants' choice did not differ from chance, 63% incorrectly chose the referred box (binomial test, $n = 32$, $p = .215$, $OR = 1.70$). In contrast to the original study, there was no significant difference in the number of infants who chose the referred box between the TB and the FB condition (Fisher's exact test, $n = 60$, $p = .293$, one-tailed, $\phi = .094$).

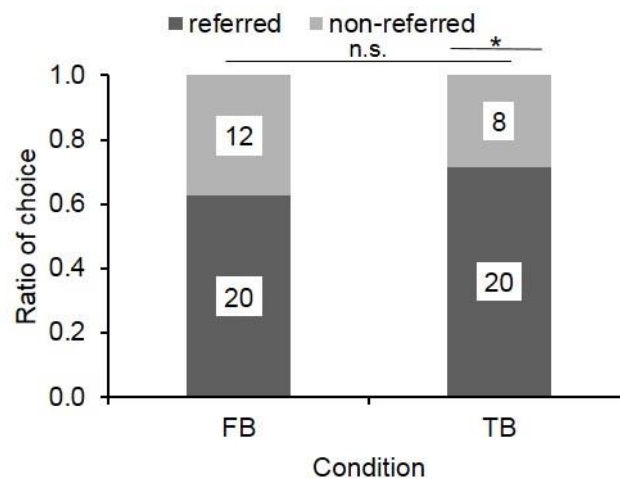


Figure 7. Interaction task. Percentage of infants who chose each box in the first trial in both conditions. Numbers in bars show number of infants. * $p < .05$

Analyses on repeated trials

48 participants took all three trials, eight took two trials, and four took one trial. When analyzing the mean performance across the repeated trials (including all participants), in the TB condition, infants chose the referred box significantly more often than expected by chance ($M = .71$, $SD = .33$; $t(27) = 3.41$, $p = .002$, $d_z = .66$, $CI: .085, .343$). However, in the FB condition, infants also chose the referred box significantly more often than expected by chance ($M = .64$, $SD = .35$; $t(31) = 2.19$, $p = .036$, $d_z = .39$, $CI: .009, .262$). Across the test trials, there was no significant difference between the conditions in infants' choice of the referred box ($t(58) = .89$, $p = .377$, $d_z = .23$, $CI: -.256, .098$).

Correlations between measures

We correlated performance in our key measures of anticipatory looking, looking time, pupil dilation, and interaction with each other. As evident from the group results, there were various ways of composing the variables for each measure (e.g. DLS, first look, number of infants, performance in a given trial or order). We report the most meaningful relations between variables for this type of exploratory analyses. No other correlations were significant. Table 1 provides a summary of the main results. Sample sizes of tests vary because only infants who provided data for both measures that were compared could be considered.

Anticipatory looking

In the Anticipatory Looking task, the first look and the DLS of the FB1 condition did not correlate with the anticipatory looking measures of the A + O task in the FB condition; did not correlate with any of the FB looking time measures of the A + O task during the outcome phase; did not correlate with any of the FB pupil size measures of the outcome phase; and did not correlate with any of the interaction measures in the FB condition. Also, none of the measures correlated with the FB2 condition. This was also true when only considering those infants that passed the second familiarization trial of the AL task.

For the anticipatory looking in the A + O task, the first look measure revealed no correlations. The mean DLS across both trials correlated with the looking time difference between incongruent and congruent outcomes. This correlation was only present in the TB condition ($r(25) = .506$, $p = .010$), and absent in the FB condition. In the TB condition, it was

present only in the second trial (in which the group performed better; $r(25) = .413, p = .040$), but not in the first trial. In the second trial of the TB condition, this correlation emerged also on the level of individual infants ($\phi(25) = .345, p = .085$). There were no relations to the Interaction task.

Table 1.

Overview of correlations between the different measures: DLS for anticipation (1) in the FB1 condition and (2) in the second trial of the A + O task, (3) looking time difference of incongruent minus congruent trials in the outcome phase, (4) the relative change in pupil dilation in incongruent trials, and (5) the mean performance over the repeated interaction trials.

Measures	1	2	3	4	5
1. Anticipation FB1	-				
2. Anticipation A+O	.036 n=21	-			
3. Looking time	-.251 n=21	TB: .413* n=25 FB: .009 n=25	-		
4. Pupil dilation	.105 n=21	.027 n=50	.711**¹ n=22	-	
5. Interaction	.175 n=24	-.272 n=44	.004 n=46	TB: .677** n=22 FB: -.039 n=23	-

* $p < .05$; ** $p < .001$; ¹for the mean pupil size, not the RCD

Looking time

In addition to the reported correlation with the anticipatory looking in the A + O task, the looking time difference between incongruent and congruent outcomes correlated with the mean pupil size during the incongruent outcome ($r(52) = .382, p = .005$). The correlation was also found when correlating the looking time difference with the pupil size in the fourth second ($r(50) = .319, p = .024$; see our additional analyses, Appendix 2). Because infants who saw the incongruent trial first seemed to perform better than the other infants, we ran the correlation for only that group of infants. This yielded an even stronger correlation for the looking time difference and the mean pupil size in incongruent trials ($r(22) = .711, p < .001$). This correlation was also significant for each condition separately (FB: $r(10) = .749, p = .013$; TB: $r(12) = .700, p = .011$) and with the pupil size in the fourth second of the reaching phase ($r(22) = .504, p = .017$). None of the other looking time measures yielded any correlations.

Pupil dilation

In addition to the reported correlation with looking times, there was a correlation between the RCD in the incongruent trial and the mean number of correct interaction trials, but only in the TB condition ($r(22) = .677, p = .001$). This correlation held when only analyzing the first interaction trial of the TB condition ($r(22) = .578, p = .005$).

Discussion

The rationale of the present study was to test for the robustness and replicability of implicit ToM tasks used in infants and toddlers, as well as for their convergent validity. To this end, we collected four implicit measures across three different infant ToM tasks, testing visual anticipation of belief-congruent actions, visual responses to violations of expected belief-congruent actions, and interactive reactions to belief-congruent requests.

Reliability of implicit ToM measures

Regarding replicability, the study revealed a mixed pattern of findings. The fact that we could replicate the pattern of findings in the true belief conditions and the FB1 condition of the AL task suggests that the measures were in principle sensitive to infants' processing of the situation and that there were no general problems with the task setups. However, none of the measures provided strong and conclusive evidence for false belief processing and, in a strict sense, failed to replicate previous findings. At the same time, our additional analyses on repeated trials, order effects and our new measure of pupil dilation indicate some degree of false belief processing under certain conditions. These indications pertained to the looking time and pupil size measures, but were absent for the anticipatory looking and interaction measures.

Regarding the looking time measure, the first trial analysis of our new VoE task which we conducted analogously to single trial analyses of previous studies (e.g., Onishi & Baillargeon, 2005), yielded only a weak effect in the TB condition, and no effect in the FB condition, thus failing to provide robust evidence for belief-based processing. This might call into question the robustness and size of single trial effects, and should be taken into consideration when designing future studies. Because we did a conceptual replication of the VoE paradigm, we do not interpret our findings as a failure of a direct replication. Note also that the original studies coded looking

times live from video while we relied on automated eye-tracking recordings. Our analyses on tracking ratios, however, revealed very good tracking and no firm grounds to question its reliability (see participants section and Appendix 2). In light of the findings of the diverse and procedurally very dissimilar VoE studies in this field one could have expected better performance (e.g., Scott et al., 2012; Surian et al., 2007; Träuble et al., 2010). However, when looking at our full-factorial analyses, the looking time measure did reveal sensitivity to belief-congruent action outcomes for true and false beliefs – but only if infants watched the incongruent outcome first. This kind of order effect is common in visual habituation research when congruent and incongruent outcomes alternate in a test phase following habituation (e.g., Baillargeon, 1987; Baillargeon, Spelke, & Wasserman, 1985). Given that our participants were already 2-years-old, it is plausible that they kept track of what they had seen in the first trial. The effect was strong enough to remain when measured at the level of individual children. The underlying mechanisms of this measure are unclear, but our descriptive findings on the temporal unfolding suggest that in the case of belief-congruent outcomes, infants begin to divert their attention fairly early, already after the first few seconds of a still frame, while in the case of belief-incongruent outcomes they keep looking for a long time. One possible interpretation is that infants wait for a congruent closure and expect a further step in the sequence and hence keep watching. If this was the case, this effect should reduce across repeated trials with incongruent outcomes.

Regarding the pupil size measure, the prediction was that a belief-incongruent outcome would yield heightened attention compared to an expected outcome, resulting in a larger increase in pupil size. Indeed, our measure of relative pupil size increase revealed sensitivity to belief-congruent action outcomes for true and false beliefs, and, like the looking duration measure, this was only the case for infants who had first watched an incongruent outcome. While the measure was not sensitive enough to reveal the effect on the level of individual children, the effect remained in a first trial analysis. Given that this is the first report of a belief-congruency processing effect in pupil size change we need to be cautious with its interpretation. The temporal dynamics of the effect show that it emerged early (after 2-3 seconds into the unfolding of the reaching event) and lasted astonishingly long. The similarity to the looking time measure suggests similar attentional processes underlying 2-year-olds looking time and pupil size change to violations of expected action outcomes. This is in line with recent findings showing that both pupil dilation and looking times increased in response to stimuli violations, for example in object

permanence tasks (Jackson & Sirois, 2009) and in face processing tasks (Falck-Ytter, 2008; Gredebäck, Eriksson, Schmitow, Laeng, & Stenberg, 2012).

Because the A + O task was always administered last, it could be that fatigue or other carry-over effects influenced performance. Previous VoE studies, however, have used multiple trial procedures, long sessions, within-designs or even included several FB tasks, and found no effects of trial or task order (Poulin-Dubois & Yott, 2018; Träuble et al., 2010; Yott & Poulin-Dubois, 2016). At any rate, fatigue should not be selective to a specific condition but affect processing on a general level. However, this is not what we found. Our study revealed selective significant processing differences in the TB condition and under specific circumstances (order) in the FB condition. Note also that the rate of tracking loss did not differ drastically between first and last task, suggesting that infants' visual attention was not influenced by potential fatigue effects. It remains possible, of course, that for currently unknown reasons false belief processing is demonstrable only in single task studies using single trials.

Regarding infants' visual anticipations, one interpretation of the pattern of results across tasks is that anticipatory looking reveals neither belief tracking nor goal anticipation but rather experience-dependent anticipations which may change within a task (Paulus et al., 2011). Findings from the familiarization trials question the suitability of visual saccades in revealing action anticipation: Although 2-year-olds have a robust understanding of reaching as goal-directed act, they anticipated in the AL task correctly only after three trials (two warm-up and one familiarization), and were still far from ceiling. The negative correlations between the first and second trials of the familiarization phases indicate some form of perseveration such that infants tended to look where they had last seen the hand appear, suggesting extraneous task demands.

Findings from the false belief conditions strongly speak against an understanding of false beliefs. The replication of the Southgate et al. (2007) results failed for the FB2 condition, which is the crucial condition for crediting participants with false belief processing. To succeed in FB1 infants just had to look at where the ball had been last. This same strategy led to failure in FB2 where the last location of the ball was the belief-incongruent location. Importantly, we did not just fail to replicate the conventional p -value, but the patterns in the FB2 condition were in the opposite direction of the original study, and they were significantly different from the FB1 condition, rejecting the hypothesis that common FB processing underlies these two FB conditions. Our A + O task provided converging negative evidence. Although the task may

perhaps exert higher demands because the agent disappears (Rubio-Fernández & Geurts, 2013) and the object remains present (Wang & Leslie, 2016), it is important to note that the task matched the verbal standard false belief task conceptually most closely.

The interaction-based measure of belief-congruent reacting, finally, did not reveal clear-cut false-belief understanding either. The replication of results by Southgate et al. (2010) failed because infants performed at chance in the FB condition (and across trials even in the opposite direction). Although infants were above chance in the TB condition, one could have expected better performance given the findings by Southgate et al. and the generally fairly easy structure of the task for 2-year-olds. Because we kept the element of deception in the TB condition (up to the point when E1 came back and watched) for better comparison with the FB condition, it is theoretically possible that this cue detracted from E1's actual epistemic state. However, even if the cue perhaps underestimated infants' above chance performance in the TB condition, the problem is that infants clearly failed in the FB condition. Why infants performed poor in the FB condition remains to be explained. Informal observations revealed that infants often made a second offer (not reported by Southgate et al.), which perhaps indicates that infants did not fully understand the referential specificity of the request.

It is at least theoretically also possible that our participants failed because they were too old for this task, since the original study tested 17-month-olds. However, this seems very unlikely, because other recent replication studies have also failed to replicate interaction-based tasks with 18-month-old infants (Poulin-Dubois & Yott, 2018), and a study using a modified version of the Southgate et al. (2010) task has recently reported positive evidence at 36 months (Király et al., 2016). It could also be that the preceding AL task influenced performance in the Interaction task. However, this too appears unlikely, because all infants engaged in the warm-up trials and enjoyed interacting with the experimenter across all trials. Further, any task order effect should not be selective but influence performance in true and false belief conditions equally. Again, this is not what we found.

Convergent validity of implicit ToM measures

Children's verbal ToM at 4 years of age is characterized by a broad and systematic competence that emerges in synchronized and correlated fashion across various superficially dissimilar tasks and measures (Astington & Gopnik, 1988; Perner & Roessler, 2012). Our correlation analyses did not reveal a comparable level of broad ToM competence for infants.

Correlations were mostly absent and scattered. This is in line with other recent studies that have failed to find unified performances between and within different FB paradigms (Poulin-Dubois & Yott, 2018; Yott & Poulin-Dubois, 2016). Bearing in mind our exploratory approach, two main findings emerged. First, the few correlations were mostly selective to the true belief condition. This indicates that the tasks did not measure totally different aspects, or suffered from very different extraneous task demands. It also underscores the interpretation that infants do not have a unified concept of belief. The TB correlation between the looking time difference and the DLS of the second A + O anticipation trial indicates convergent validity and supports the interpretation that by the second test trial infants had learnt to correctly anticipate (but only in the TB condition). The TB correlation between the pupil size increase in incongruent trials and the Interaction task is interesting as it may reflect as common denominator between the two variables a concern for others (Hepach, Vaish, & Tomasello, 2012). Again, however, the correlations were selective to the TB condition and hence do not indicate an understanding of false beliefs.

The second finding was the relation between the looking time and the pupil size. It is the only correlation that holds for both true and false beliefs, which provides convergent validity to the positive group level results of false belief understanding in these two measures. Confirmatory results are required, especially because the correlation concerned the absolute tonic size of the pupil, not the relative increase of the pupil. However, it is unlikely that the finding reflects only a peripheral physiological correlate of the looking time pattern (longer looking leads to larger pupil), because the correlation was already evident when calculating with the fourth second of the reaching phase. Our additional control analyses (see Appendix 2) further speak against measurement artifacts.

The current study revealed mixed findings depending on measures and analyses. The two attentional measures, looking time and pupil change, were most promising in revealing false belief processing. This could be because they were least demanding and rather unconscious, at least in the case of pupil change (which cannot be produced consciously). Anticipatory looking and interactive behaviors are more goal-driven than looking time and pupil change. They entail anticipating specific consequences of the behavior (e.g., to see an event in a specific location; to satisfy a requester's need), while looking time and pupil change rather retrodict than project events. Anticipatory looking and interactive behaviors entail elements of choice (a decision where to look, or what to offer), which perhaps requires more reasoning processes than for the

other two measures. Therefore, and because of their anticipatory direction, anticipatory looking and interactive behaviors depend on pragmatically very clear situations. AL, and especially interaction tasks, however, often suffer from weak pragmatic soundness. For example, in the video-based tasks, the agent was only weakly introduced to reach reliably correctly – a necessity to anticipate a correct reach; and in the Interaction task, the requester ultimately may have wanted both toys. Further, it was not conveyed why he did not retrieve the toys himself. Because the pragmatics of these non-verbal tasks are often difficult to convey, these measures may be more difficult to replicate. It remains debatable, however, how to best account for the level of ToM understanding exhibited in the looking time and pupil change measures given that there were no comparable indications of such an understanding in the other tasks.

Conclusion

From a theoretical point of view, the pattern of findings in the present study challenges strong claims about infant ToM, in particular nativist and two-systems views. Despite fundamental disagreements in some respects, these two kinds of accounts converge on the claim that different implicit tasks all measure the same basic ToM capacity (ToM proper, according to nativist accounts; a basic, efficient and automatic ToM system, according to the two-systems-view). These tasks should thus each by itself be robust and there should be convergence and correlation across them.

The present study fails to find robust evidence for either replicability or convergent validity of the implicit ToM tasks. Now, what does such absence of evidence amount to? On the one hand, the limited support from the VoE task, given that this was a conceptual rather than direct replication, with many differences between original and replication task, by itself cannot be considered evidence for the absence of robust replicability. On the other hand, however, two of the three tasks used here (AL and interaction) involved direct replication attempts of previous studies, and the clearly negative present findings can thus constitute at least *prima facie* evidence of absence of robust replicability of these tasks. But clearly, the present findings taken by themselves cannot settle the broader question of the robustness and replicability of implicit ToM tasks. In order to understand whether implicit ToM in infants and toddlers is a real and robust phenomenon, future research will need to design and administer systematic, large-scale, pre-registered multi-lab replication and meta-analytical studies.

In the meantime, the present findings (together with other convergent evidence reported in this Special Issue) suggest that the empirical foundation for positing early and implicit ToM competence in infants is much less robust and conclusive than previously assumed.

Acknowledgements

This study was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft), research unit “Crossing the borders: The interplay of language, cognition, and the brain in early human development” (Project: FOR 2253, Grants: LI 1989/3-1 and RA 2155/4-1). We want to thank Victoria Southgate for sharing her video material and giving advice on her tasks. We also want to thank the colleagues and students who helped in data collection and recruiting participants, especially Marianna Jartó, Wiebke Pätzold, Nicola Ballhausen, Betty Timmermann, Maria Häberlein, Maximilian Kaliski, Jula Brüning-Wessels and Jana Klose.

Appendix 1. Stimuli & Procedure

Anticipatory Looking task

We showed infants two warm-up trials, two familiarization trials and two false belief test trials. One test trial was in the false belief 1 (FB1) condition, the other test trial in the false belief 2 (FB2) condition. In all trials, an agent stood behind a screen that contained two doors, each in front of one box. The agent wore a visor cap hiding her eyes, while she was following the displayed actions with her head movements. In the final phase of each trial, the two doors were illuminated for 1s which was accompanied by a chime sound. This was followed by a 1.75s still frame (without illumination), after which the agent would reach through one of the doors. The warm-up and familiarization trials served to increase infants’ understanding of the actor’s goal and the predictability that the agent will reach through the door after the illumination and still frame phase.

In the warm-up trials (Senju et al., 2009), infants saw a red whale toy sitting on one box (one trial on each box, order counterbalanced between subjects). After the illumination and the still frame, the agent reached through the door for the toy. In the two familiarization trials, the agent watched a teddy putting a toy in a box (one trial on each side, order counterbalanced

between subjects) and afterwards leaving the scene. After the illumination and the still frame, the agent reached through the door into the box that contained the toy (only in the first trial she also took it out).

In a FB1 test trial, the teddy put the toy in a box but suddenly decided to change the toy to the other box and to leave the scene (note: this was all seen by the agent). Afterwards, the agent was distracted by a phone call and turned to the back. Thus, she did not witness how the teddy reappeared, took the toy out of the box and left the scene giggling with the toy. In a FB2 test trial, the teddy would leave the scene after he put the toy in the first box. While he was gone, the agent got a phone call and turned to the back. So, contrary to the FB1 condition, in the FB2 condition the agent was distracted and did not witness how the teddy reappeared and changed the ball to the other box. Also in the FB2 condition, the teddy decided to take the toy out of the second box again and to leave the scene with the toy (unseen by the agent). After the illumination and the still frame, infants had to predict through which door the agent would reach in order to search for the toy. No outcome was shown in the test trials. As in the original study, the last object location was always the same in both conditions but we counterbalanced the side between subjects.

Anticipation + Outcome task

We showed each infant two familiarization trials and two test trials. We counterbalanced the order for target side and congruency. An agent (same agent as E1 in Interaction task) sat behind a screen that contained two doors with two boxes in front. He followed the actions of a teddy bear with his head movements. All trials started similarly, the agent noticed a ball that was already positioned centrally between the boxes and he said, "Ah." A teddy appeared centrally from the bottom of the scene and waved to the infant and the agent, while the agent said, "Hello." The teddy opened the lid of the first box, put the ball inside and closed the lid. In a familiarization trial, the teddy waved goodbye and left the scene centrally to the bottom. Then, the agent said, "Okay," and ducked down behind the screen, so that he was not visible anymore. Subsequently, the two doors were illuminated for 1s which was accompanied by a chime sound. After a delay of another 2s, a fixation cross appeared centrally between the doors for 550ms, which served to center infants' gaze before the outcome. Afterwards, the agent reached through the door inside the baited box (being visible through the door). He grabbed the ball and

reappeared above the screen. He held the ball in his hand, smiled and said, “Ah,” alternating his gaze between the ball and the infant.

In the test trials, after the teddy put the ball in the first box, he opened that box again, placed the ball between the boxes, opened the second box and put the ball inside. He closed the lid of the second box and afterwards the lid of the first box (the teddy always closed the lid of the new toy position first, so that the agent’s last view went to the empty box). Subsequently, the teddy waved goodbye and left the scene centrally to the bottom. A phone call sound was played, which was commented by the agent with, “Oh, telephone,” while he faced the infant. The agent stood up, turned around and disappeared centrally through a gap in between the two black walls in the background. In the true belief condition, the agent would reappear after a delay of 3s and, thus, he would witness all subsequent events. In the false belief condition, on the contrary, the agent was still gone while the teddy reappeared, opened the baited box, placed the ball in between the boxes, opened the other box, placed the ball inside and closed the lids in the former manner. The teddy disappeared right before the agent in the false belief condition reappeared from behind the background (note: the ball was still present). The agent sat down behind the screen and from now on all events happened parallel between the two test conditions. As in a familiarization trial, the agent said, “Okay,” and ducked down behind the screen, so that he was not visible anymore. The two doors were illuminated for 1s which was accompanied by a chime sound. After a delay of another 2s, a fixation cross appeared centrally between the doors for 550ms. Afterwards, infants were shown an outcome phase that consisted of a reaching phase (7s) in which the agent reached through the door inside the baited box (belief-congruent in TB, but incongruent in FB) or the empty box (belief-congruent in FB, but incongruent in TB), followed by a still frame phase (20s) in which infants were shown a still frame of the agent with his hand inside of the box.

Interaction task

Before the test trials, infants were presented with two warm-up trials designed to familiarize them with searching for objects in the boxes. The experimenter (E1) gave the infant two familiar objects and allowed it to explore them for about 10s. Afterwards, E1 put the objects into the two boxes and closed the lids. Then he asked the infant to bring him one of the objects by naming it. If the infant succeeded by bringing the correct toy first, the child was asked for the

other object. This was repeated until the infant brought the requested object in two consecutive trials.

In the three test trials, E1 showed the infant two novel toys and allowed it to explore them for about 10s, placed them in the two boxes and closed the lids (objects were not labelled yet). E1 told the infants that he had to go out because he forgot something but he would be back soon. E1 left the room through the door and another experimenter (E2) that was unknown to the infants entered the room from behind the curtains. E2 emphasized her deceptive plan by giggling and gesturing, “Shush.” E2 sat down between the boxes and interchanged the objects. E2 opened both boxes, placed one object in front of its box, took the other object, showed it to the infant, placed in the other box, picked up the first object, showed it to the infant, placed it in the other box, and closed both boxes simultaneously. In the TB condition, E1 would reappear as “early bird” in the moment when E2 had opened the boxes and placed the first object in front of the box, but right before she would start interchanging the objects and he would conspicuously watch E2. From the moment when E1 entered the room in the TB condition, E2 stopped acting deceptive. In the FB condition, however, E1 would re-enter the room shortly after E2 hid behind the curtains again. E1 sat down between the boxes in a position from where he could not look inside and asked the infant, “Do you remember what I put in here? There is a Sefo in here. Shall we play with the Sefo? Can you give the Sefo to me?” whilst pointing to one of the two boxes. E1 opened both boxes simultaneously and faced the infant. E1 asked repeatedly for the objects until the infant began to point to or to approach one of the boxes. We counterbalanced target side, object pair and target object between the trials for this task.

Appendix 2. Results

Analyses on looking time in A + O task including participants with more weighted gaze samples

We repeated our main analysis on the looking time measure but only included participants with higher amounts of gaze samples (>50%). The interaction between congruency and order remained significant up to the point when only participants with gaze samples >80% were included (see Table A1). For all configurations, simple comparisons revealed that infants looked significantly longer in the incongruent trial compared to the congruent trial, only if the incongruent trial was administered first (all $ps < .05$); and there was no difference between both

trials when the congruent trial was first. The effect vanished at >90% gaze samples and only 22 participants left.

Table A1.

Main results of the 2x2x2 ANOVA with congruency as within subject factor and condition and order as between subject factors on the looking time during the outcome phase of the A + O task by only including participants with more weighted gaze samples (%WGS) and the resulting sample sizes (N).

%WGS/Factors	congruency	congruency*order	condition	order	N
>50	$F(1, 44) = 6.1, p = .017,$ $\eta_p^2 = .122$	$F(1, 44) = 7.2, p = .010,$ $\eta_p^2 = .141$	n.s.	n.s.	48
>60	$F(1, 41) = 5.1, p = .030,$ $\eta_p^2 = .110$	$F(1, 41) = 7.5, p = .009,$ $\eta_p^2 = .154$	n.s.	n.s.	45
>70	n.s.	$F(1, 36) = 5.6, p = .023,$ $\eta_p^2 = .135$	n.s.	n.s.	40
>80	n.s.	$F(1, 28) = 4.6, p = .040,$ $\eta_p^2 = .142$	n.s.	n.s.	32
>90	n.s.	n.s.	n.s.	n.s.	22

Mirrored vs. un-mirrored videos in the AL task

To exclude that the mirroring of the test videos had an effect on infants' performances in the AL task, we compared performances of those infants who passed the second familiarization between mirrored and un-mirrored videos. There was no difference between performances for the first look measure in both conditions (Fisher's exact test; FB1: $n = 28, p = .375$; FB2: $n = 27, p = 1.0$) and no difference for the DLS measure (FB1: $t(27) = 1.63, p = .114$; FB2: $t(27) = .410, p = .685$).

Analyses on distance to the eye-tracker

To rule out that a variable distance between the eye and the eye-tracker could explain findings of pupil dilation during the outcome phase of the A + O task, we conducted a 2x2x2 ANOVA for a relative change in distance (baseline was the last second of the anticipation phase; the focal phases were the mean distance over the whole outcome phase and the distance in the fourth second of the outcome phase) with congruency as within subject factor and condition and order as between subject factors. There were no main effects for congruency (whole outcome: $F(1, 45) = .10, p = .752, \eta_p^2 = .002$; fourth second: $F(1, 44) = .02, p = .888, \eta_p^2 = .000$) or condition (whole outcome: $F(1, 45) = 1.83, p = .182, \eta_p^2 = .039$; fourth second: $F(1, 44) = .83, p = .369, \eta_p^2 =$

.018) and no interaction between congruency and order (whole outcome: $F(1, 45) = .00, p = .960, \eta_p^2 = .000$; fourth second: $F(1, 44) = .06, p = .812, \eta_p^2 = .001$). Further, pupil size and distance to the eye-tracker were not correlated in congruent (whole outcome: $r(52) = -.099, p = .484$; fourth second: $r(51) = -.190, p = .183$) or incongruent trials (whole outcome: $r(52) = -.060, p = .673$; fourth second: $r(51) = -.034, p = .811$). According to this, a variable distance to the eye-tracker cannot explain our findings on pupil dilation.

Offset illumination analyses for the first trial of the AL task

When analyzing or measurements offset of the illumination, in the second familiarization trial, 29 of 48 (60%) of the infants directed their first look to the correct door (binomial test, $p = .193, OR = 1.50$). Of those, 26 provided data for the first test trial, 11 in the FB1 condition, and 15 in the FB2 condition. When analyzing the first test trial of those infants who had looked at the correct door in the second familiarization trial, in the FB1 condition 55% of the infants directed their first look to the correct door (binomial test, $n = 11, p = .500$, one-tailed, $OR = 1.22$); and their looking time to the correct door did not differ from chance ($M = .394, SD = .809; t(10) = 1.62, p = .137, d_z = .49, CI: -.149, .937$). In the FB2 condition, infants first look went significantly more often to the incorrect door than expected by chance (13% correct; binomial test, $n = 15, p = .007, OR = 6.69$); and infants tended to look longer to the incorrect door ($M = -.368, SD = .736; t(14) = 1.94, p = .073, d_z = .50, CI: -.776, .040$). The two conditions differed significantly from each other on both measures (first look: Fisher's exact test, $n = 26, p = .038, \phi = -.441$; DLS: $t(24) = 2.50, p = .020, d_s = .98, CI: .133, 1.391$).

We additionally analyzed the total looking time to each door during the 1.75s after offset of the illumination for the first trial for those who passed the second familiarization trial. A 2x2 ANOVA with window (correct, incorrect) as within-subject factor and condition (FB1, FB2) as a between-subject factor yielded no significant main effects for window ($F(1, 24) = .20, p = .331$, one-tailed, $\eta_p^2 = .008, CI: -.251, .388$) or condition ($F(1, 24) = 1.60, p = .219, \eta_p^2 = .062, CI: -.373, .90$), but a significant interaction between condition and window ($F(1, 24) = 5.52, p = .027, \eta_p^2 = .187$). Infants tended to look longer at the correct door ($M = 632\text{ms}, SD = 601$) compared to the incorrect door ($M = 199\text{ms}, SD = 255$) in the FB1 condition ($F(1, 24) = 3.38, p = .078, \eta_p^2 = .124, CI: -.53, .918$), but look longer at the incorrect door ($M = 704\text{ms}, SD = 529$) compared to the correct door ($M = 409\text{ms}, SD = 458$) in the FB2 condition ($F(1, 24) = 2.15, p = .156, \eta_p^2 = .082, CI: -.710, .121$), though not significantly.

RCD in the fourth second of the reaching phase

To test whether the effect was immediately present at the beginning of the outcome phase, we focused on the reaching phase when the hand started approaching the container and calculated the RCD for the fourth second of the reaching phase (see Figure 7) with the last second of the anticipation phase as baseline. Findings were the same. A 2x2x2 ANOVA with congruency as within subject factor and condition and order as between subject factors revealed a significant main effect of order ($F(1, 44) = 4.50, p = .040, \eta_p^2 = .093, CI: -.052, -.001$) and an interaction between congruency and order ($F(1, 44) = 3.99, p = .052, \eta_p^2 = .083$). Infants' relative pupil increase tended to be larger in the incongruent ($M = .101, SD = .064$) compared to the congruent outcome ($M = .076, SD = .050$) when the incongruent outcome was presented first ($F(1, 44) = 3.80, p = .058, \eta_p^2 = .079, CI: -.051, .001$), but not when the congruent outcome was presented first ($F(1, 44) = .69, p = .412, \eta_p^2 = .015, CI: -.014, .034$). For each condition separately, this comparison did not reach conventional significance level.

A first trial analysis provided confirmatory support: A between subject univariate ANOVA with order and condition as factors revealed that infants in the group with the incongruent outcome had a significantly larger increase in pupil size compared to infants in the group with the congruent outcome ($F(1, 46) = 4.49, p = .039, \eta_p^2 = .089, CI: -.067, -.002$), with no significant difference between conditions and no interaction. For each condition separately, this comparison did not reach conventional significance level.

Study 2: The sefo task: A measure of early false belief understanding?

This study is a collaboration of Sebastian Dörrenberg (University of Hamburg), Lisa Wenzel (University of Göttingen), Marina Proft (University of Göttingen), Hannes Rakoczy (University of Göttingen) and Ulf Liszkowski (University of Hamburg), and is not published, yet.

Abstract

The last decade produced astonishing findings that even young infants are capable of false belief (FB) representation, the litmus test for crediting a Theory of Mind. However, a variety of recent replication studies question the reliability of original findings. In particular, a recent replication attempt of the sefo task (an interactive FB task by Southgate et al., 2010) was unsuccessful in finding FB representation in 24-month-olds (Dörrenberg et al. 2018). Another sefo replication study failed with older children (Grosse Wiesmann et al. 2017). Surprisingly, the sefo task suffers from weak pragmatic soundness. For instance: Why does the experimenter leave the room? Why does she not retrieve the toy herself? To clarify why these studies failed to reproduce the original findings, (i) we conducted close direct replications with the original age group (17-month-olds), (ii) we developed pragmatic modifications by introducing an apparatus that created a fun game with the objects, that gave the experimenter a reason to leave the room and that occupied his hands on the return, and (iii) we validated the sefo task by testing 3-year-olds in a direct replication and additionally in a standard FB task. Results were negative across tasks and age groups. Our findings question the suitability of the sefo task to measure FB understanding in young children.

Introduction

The classical view on the development of Theory of Mind (ToM) has been revised by new findings: Already young infants ascribe false beliefs (FB) to other persons (e.g., Onishi & Baillargeon, 2005). Over decades, studies with explicit measures of ToM (the ability to ascribe subjective mental states to others) found across-the-board that only from age four on, children are able to pass FB tasks, which require the understanding that someone has a misrepresentation of reality (Wellman et al., 2001; Wimmer & Perner, 1983). However, a variety of new implicit

measures of FB understanding suggest that extraneous demands of explicit tasks (likely linguistic and inhibitory) camouflaged young children's competence (see e.g., Scott & Baillargeon, 2017). For instance, by using non-verbal measures and otherwise simplified tasks, even very young infants correctly anticipate an agent's action who is mistaken about an object location in anticipatory looking tasks (Clements & Perner, 1994; Southgate et al., 2007), they are surprised when an agent acts contrary to his FB in violation-of-expectation tasks (Onishi & Baillargeon, 2005; Träuble et al., 2010), and they offer appropriate helping behavior for agents in interactive FB tasks (D. Buttelmann et al., 2009; Southgate et al., 2010). Although some argue these findings on implicit tasks could be explained in more parsimonious ways than true FB understanding, e.g. by applying behavior rules or by reacting to perceptual novelty (Heyes, 2014a; Perner & Ruffman, 2005), also far-reaching theoretical accounts emerged, such as two-systems accounts that suggest the existence of two mindreading systems, i.e. an early efficient and a later-emerging flexible ToM system (Apperly & Butterfill, 2009; Low et al., 2016).

Unfortunately, as in other areas of psychological science (e.g., Open Science Collaboration, 2015), infant ToM research currently faces a replication crisis (for discussion, see Baillargeon, Buttelmann, & Southgate, 2018; Poulin-Dubois et al., 2018). That is, a variety of recent replication studies on various measures of FB understanding in infants yielded negative and conflicting results (e.g., Crivello & Poulin-Dubois, 2018; Dörrenberg et al., 2018; Kulke, Reiß, Krist, & Rakoczy, 2018; Powell et al., 2018; Schuwerk et al., 2018). These replication failures seriously question the reliability of the original effects and the existence of early FB understanding per se. However, especially direct replications that are conducted as closely as possible to the methods of the original study can provide conclusive evidence on the reliability of existing findings. That is because deviations from study population or experimental procedure could also account for replication failures (Baillargeon et al., 2018; Rubio-Fernández, 2018b). This might be particularly relevant for interaction-based tasks, where the test situation is influenced by the interaction style of the experimenter, and where it can be challenging for a replication study to script the test procedure from selected video recordings or from method sections that are often shortened. This study aims at taking a closer look at the replicability of the "sefo task", originally from Southgate et al. (2010). In case of this interactive FB task, there are two published replication studies that failed to find the original effects (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2017, Supplement), and one study that partially replicated (Király, Oláh, Csibra,

& Kovács, 2018). However, each of these replication studies had different methodological limitations.

In the original sefo task study, an experimenter (E1) showed two objects to the infant, put each object in a separate box and left the room. Another person (E2), who was hidden behind curtains and unknown to the infant (to emphasize her deceptive intentions), swapped the objects. This was either seen by E1 (E1 came in as early bird in the true belief (TB) condition) or not seen (E1 was still outside in the FB condition). Afterwards, E1 pointed at one box and requested the child to retrieve the object. In three experiments that differed in the phrasing of the experimenter's request, 17-month-old infants interpreted E1's communicative reference differently when she held a FB versus a TB. That is, in the FB condition infants chose the non-referred box more often, while in the TB condition they chose the referred box more often, and there was a significant difference between conditions. A pattern that suggests FB understanding.

The two non-replication studies failed to find this pattern (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2017), neither did children choose the non-referred box more often in the FB condition (the majority actually chose the referred box), nor was there a condition difference (only the study by Dörrenberg et al. conducted a TB condition). Interestingly though, the study by Grosse Wiesmann et al. found that performance of 3- and 4-year-olds in the sefo task was correlated with performance in a standard verbal FB task. This suggests that the tasks may measure the same competence (i.e., explicit ToM). But since they found only a weak correlation (.248) and only an insignificant trend, more data is required to confirm their findings. However, there were methodological differences that could explain the poor performance in these studies. Both non-replication studies tested older children compared to the 17-month-olds of the original study. In the study by Dörrenberg et al. (2018) 24-month-olds were tested, and in the study by Grosse Wiesmann et al. (2017) 3- and 4-year-olds were tested. Therefore, it possible that the replications failed because performance in the sefo task declines with age, and the task could be suited exclusively for testing infants in the second year of life. Another limitations of the study by Grosse Wiesmann et al. (2017) was that the second person, who tricked the experimenter in the FB condition by swapping the objects, was present in the experimental room the whole time and not hidden as in the original study. Other studies on verbal FB tasks have shown that deceptive behavior can enhance performance of 3-year-olds (Sullivan & Winner, 1993; Wellman et al., 2001). The presence of the tricker throughout the task may have interrupted the deceptive mode and led to weaker performance (as the authors themselves

suggest). Thus, it may be that older children pass the task when trickery is exerted similar to the original study.

The replication study by Király et al. (2018) found positive evidence for retrospective false belief attribution. When children were informed after the object swap that E's sunglasses were opaque in the FB condition but transparent in the TB condition, performance differed between conditions in 36-month-olds but not in 18-month-olds. This study also found that response of 18-month-olds differed between conditions when they learned about the opacity of the sunglasses before the object swap (prospective belief tracking). But since 18-month-olds performed only at chance in the FB condition and were tested with a different procedure compared to the original study (e.g., using sunglasses), that study does not provide conclusive evidence for the reliability of the original study either. The pattern Király et al. found constitutes a partial replication and extension of the findings by Southgate et al. (2010). However, there may also be alternative explanations for their findings. As in Experiment 1 of the original study, the experimenter's prompt for requesting the object in the study by Király et al. was, "Do you remember what I put here? I put a sefo here. Shall we play with the sefo?" This prompt provides a cue to the non-referred box (which is correct in the FB condition) without any mental state attribution. That is, when participants make use of a literal interpretation of the prompt, they only need to remember that the experimenter initially put in the referred box the object that is now in the non-referred box. To rule out this alternative explanation, Southgate et al. (2010) conducted further experiments with infants where they modified the phrasing of the prompt ("Do you know what's in here?" instead of "Do you remember what I put here?") and found similar results. Importantly though, 3-year-olds, as tested by Király et al., have much better linguistic abilities than infants and may thus be able to understand and use the experimenter's prompt to solve the task. Apparently, it is important to use an unconfounded phrasing of the prompt, especially when testing older children.

However, understanding the experimenter's pointing gesture and prompt as a request for a specific object is particularly important for taking into account her misrepresentation of the locations of the toys when making a response. An unambiguous interpretation of the pointing requires to comprehend the verbal prompt. Otherwise, younger children could understand the pointing as "Open that box" instead of "Give me what (I think) is in this box". Thus, especially for infants that lack sufficient language skills, the test procedure has to offer sufficient information to solve the task. In order to do this, the original sefo task study was based on previous work

suggesting that toddlers track other's FB to assign reference to a new object label (Carpenter, Call, & Tomasello, 2002; Happe & Loth, 2002). These findings, though, have recently been challenged by a study showing that advantages of word learning in FB tasks vanished when compared to matching control conditions (Papafragou, Fairchild, Cohen, & Friedberg, 2017). Surprisingly, apart from the word learning context, the sefo task suffers from weak pragmatic soundness. For instance, after E puts the toys in the boxes, she leaves the room for no reason. It may not be clear to infants that the "game" is ongoing and that she still cares about the objects on her return. Also, when E returns, she sits right next to the boxes and could easily reach the toys herself, but she asks the infant to do so for no obvious reason. When the infant offers the toy to E, there is no outcome, nothing that would explain why she needed the toy in the first place. Uncertainty about E's intentions may explain why infants often offered the second object from the other box too after making an initial choice in the original study (personal communication with V. Southgate) and in the replication study by Dörrenberg et al. (2018). Thus, it is possible that infants would perform better in the sefo task when conducting an ecologically valid and unambiguous test procedure.

The current study was conducted to establish clarity on the reliability of the sefo task through a multi-lab replication approach. In order to overcome limitations of previous replication studies due to methodological differences, we conducted direct replications where we aimed at being as accurately as possible on the original set-up and procedure (e.g., using similar objects and a similar mode of trickery). First, we tested participants at 17 months of age to clarify whether the task works in the age group of the original study, which has not been done yet. Second, we tested participants at 3 years of age to evaluate whether this task is indeed not suited for testing older children. To rule out alternative explanations for success in the sefo task, such as a confounded prompt (as in the study by Király et al. 2018, "Do you remember what I put in here?"), we used a phrasing of the experimenter's prompt that did not refer to the child's memory (as in Southgate et al. 2010, Experiment 3, "Do you know what's in here?"). In addition, in order to confirm that the sefo task measures the same competence as standard verbal FB tasks (Grosse Wiesmann et al., 2017), we administered 3-year-olds a standard change-of-location task and tested for correlations. Third, to test whether pragmatic shortcomings of the sefo task could make it difficult for young children to pass, we designed a pragmatically modified task version and introduced an apparatus in the procedure, which resolved several issues: The experimenter had to leave the room to get the apparatus (while toys are being swapped in his absence), it

occupied his hands when he returned (so he could not retrieve the toys himself but needed help from the child), and it created a fun game with the toys (therefore they needed a specific toy). To further assess whether participants would learn to solve the task during the interaction, we administered the 17-month-olds and the participants in the modified sefo task multiple test trials and offered them helpful feedback after incorrect trials.

Methods

Direct replication

Participants

The final sample consisted of 48 17-month-olds (twice the sample size of each experiment of the original study; median age = 17 months; 14 days, age range = 16;11 – 18;9, 25 girls and 23 boys) and 48 3-year-olds (median age = 41.5 months, age range = 36 – 47 months, 27 girls and 21 boys). Each half of the sample was tested in Göttingen and Hamburg by different experimenters (the TB condition of the 3-year-olds was tested only in Göttingen). Participants were recruited from a databank of children whose parents had previously agreed to participate in child studies. 26 more children were tested but excluded because they refused to participate (17mo: 14, 3yo: 1), failed the warm-up trials (17mo: 9), parent error (17mo: 1) or experimenter error (3yo: 1).

Design and Procedure

17-month-old participants received three test trials in the same condition (TB or FB) with helpful feedback after incorrect trials. We administered all 3-year-olds a single test trial of the sefo task, and additionally one trial of a standard verbal FB task (task order counterbalanced). Half of the children of each age group received FB and the other half TB scenarios (between-subject).

Sefo task

Children were seated on the floor between their parent's legs. Two boxes (lids attached so they remained in an upward position when opened) were positioned 120cm from the infant

and 100cm apart facing the child. Experimenter 2 (E2) hid behind white curtains. Before the test trials, at least two warm-up trials were conducted: Participants were allowed to explore two familiar objects (a bathing dug and a toy shovel) for about 10s. Afterwards, experimenter 1 (E1) put one object in each box and closed the lids. Then E1 asked to bring one of the objects. This was repeated until the participant brought each object once in two consecutive trials (original inclusion criterion). Note, some 17-month-olds had trouble doing the warm-up. Thus, to reduce the number of drop-outs, we had to modify the warm-up procedure slightly for some participants (e.g., leaving the boxes open when requesting an object) and applied a more lenient passing criterion (bringing both objects needed not to be in consecutive trials).

In the test trials, E1 presented two novel objects (a water can spout and a lemon squeezer in the first trial for the 17-month-olds as in the original study, and new object pairs for the other trials; a curtain holder and a plant watering bulb for the 3-year-olds (objects adjusted to older age); see Figure 8 for objects), allowed to explore them for about 10s, placed each in a box and closed the lids. Then E1 left the room and E2 that was unknown and hid behind curtains entered. E2 sat down between the boxes and swapped the objects: E2 opened both boxes, placed one object in front of its box, took the other object, showed it to the infant, placed in the other box, picked up the first object, showed it to the infant, placed it in the other box, and closed both boxes simultaneously. E2 emphasized her deceptive plan by whispering, giggling and gesturing, “Shush.” In the TB condition, E1 would reappear before E2 started to swap the object and would thus witness all events. In the FB condition, however, E1 re-entered the room shortly after E2 hid behind the curtains again. E1 then sat down between the boxes in a position from where she could not look inside and asked the infant, “Do you know what’s in here? I want to play with this!” whilst tapping at one of the two boxes (side and target object counterbalanced). E1 opened both boxes simultaneously, faced the infant and asked, “Can you give it to me?” E1 repeatedly requested the object until participants pointed at or approached a box. We coded the box that participants approached or pointed at first, which was either the referred box (correct in TB, incorrect in FB) or the non-referred box (incorrect in TB, correct in FB).

After offering E1 an object, the 17-month-olds received helpful feedback. When they chose correctly, E acted surprised about the new location in the FB condition but happy about the correct toy. When they chose incorrectly, E1 acted surprised and said, “Humph. Strange. No, that is not what I meant.” She then looked in the other box and said, “Ah. That is what I meant. How did it get here?”

Standard belief task

In the standard change-of-location task, the participants were positioned as in the sefo task (but two containers were 50cm apart and 50cm away from child) and E1 acted the story of a cuddly toy lynx (Luchsi) on the floor. Luchsi showed his toy car to the child and played with it briefly. Then he put his car into one of two little boxes and left the scene. In the absence of Luchsi (FB), or after his return (TB), an ape puppet appeared, swapped the car to the other box and left. E1 then asked three control questions (“Where did Luchsi put his car in the beginning?”, “Where is the car now?” and “Who put it there?”). Children got corrective feedback when answering incorrectly on the control questions, thus no child was excluded due to failure. Then, E1 asked an explicit test question (“When Luchsi returns, where will Luchsi look for his car first?”). Children could either indicate that Luchsi will search in the box containing the object (correct in TB, incorrect in FB) or in the empty box (incorrect in TB, correct in FB).

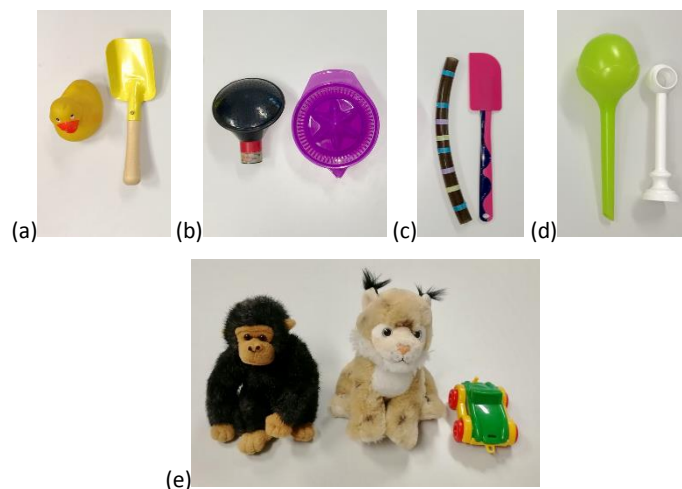


Figure 8. Objects used in the direct sefo task replication (a – d; (a) for familiarization trials, (b) for first trial of 17-month-olds, (c) and (d) for second and third trial of 17-month-olds (counterbalanced), (d) for 3-year-olds) and in the standard belief task (e).

Pragmatically modified task

Participants

36 24-month-olds (median age = 24;20, age range = 23;30 – 25;12, 17 girls and 19 boys) were tested in the metropolitan city Hamburg, recruited from a databank of children whose

parents had previously agreed to participate in infant studies. Two more infants were tested but excluded, because they refused to participate.

Materials

We used a new set of materials for each trial, each consisted of: two boxes that each contained different objects (boxes always contained three objects of the same kind), and an apparatus that could be used with the objects from the boxes for a fun game. In the familiarization trials, only the objects from one of the boxes worked with the apparatus. In the test trials, objects from both boxes would potentially work with the apparatus, though it was not obvious which object type would be the match. Figure 9 depicts all materials. In one familiarization trial, the apparatus was (a) an open plastic bottle that contained a chime and was installed diagonally through a box, which could be used with the marbles from one box (running down the bottle, eliciting the chime) but not with the cloths from the other box. In the other familiarization trial, it was (b) a ball run where the balls from one box could run down but the bricks from the other box would not work. For the three test trials, we used three different sets: (c) bellows that could be used to shoot either purple paper shucks or blue pieces of sponge, (d) an upright tube on a board with a rattle on the bottom where chestnuts could be thrown in to elicit the rattle or rattling plastic rings could be thrown over, and (e) a slingshot that could shoot either wine corks or clothespins into an attached plastic bottle. Order of material sets and target objects was counterbalanced.

Design and Procedure

Each 12 participants were assigned to one of three conditions (between-subject): TB, FB feedback, or FB no feedback. The set-up was similar as in the original sefo task, but boxes remained open throughout a trial after exploring the content (lids lying behind). Each session started with two familiarization trials, followed by three test trials in the same condition. Participants learned in the familiarization that E1 leaves the room to get an apparatus that is needed for a game with the objects from the boxes, that he then indicates which of the two object types they need for the game, and that the other object type would not work. This should ensure that in the test trials infants would understand the specificity of E1's request in the test trials.

First, E1 and the child explored the content of the two boxes. E1 then said, "I know what we can play with this. I go out and get something for us." Before E1 left the room, he checked the content of each box again and said, "Okay, that is in here, and that is in here." In familiarization trials, E1 returned with an apparatus, slowly walked towards the boxes (centered between the boxes), pointed at one box (acting as if it was struggling to hold the apparatus, stressing the need for help) and said, "We need the [toy name; e.g., the balls]. Can you give it to me?" After infants gave E1 the correct toys, they used them to play with the apparatus (e.g., roll a ball down the ball run for three times). E1 then showed the infant that the other object type would not work with the apparatus (e.g., the bricks would not roll down).

The test trials were similar as the familiarization trials, but E2 entered the room from behind the curtains after E1 left. E2 looked at the infant said, "Hello" and looked at the door to ensure that E1 was absent. In the TB condition, E1 would now re-enter the room with the apparatus in his hands, greet E2, and position beside the set-up to observe E2. In the FB conditions, E1 was still absent at this point and E2 acted in a sneaky manner (e.g., whispering, giggling, looking at the door). E2 went between the boxes, took out the objects from one box and placed them in front of the box, then she took out the objects from the other box and showed them to the child (and E1 in TB condition), then she swapped the objects from one box to the other and said, "Look, I put this here." In the TB condition, E1 commented, "Ah, okay." E2 then said goodbye and disappeared behind the curtains again. Subsequently, E1 re-entered the room through the door with the apparatus in his hands in the FB conditions, or positioned behind the boxes in the TB condition, respectively. E1 slowly walked towards the boxes (centered between the boxes), pointed at one box (side counterbalanced) and said, "We need what is in there. Can you give it to me?" He stopped behind the boxes and waited until the infant made a choice.

In the FB no feedback condition and in the TB condition, right after the infant gave an object to E1 (correct or incorrect) someone knocked at the door, E1 went to the door and acted as if he would be talking to someone. When he returned to the child, he put away the materials, claimed to have even better toys, and started a new trial. In the FB feedback condition, if the child correctly offered the objects from the non-referred box, E1 acted surprised about the new location, but they played with the toys (e.g., shooting the three sponges with the bellows). If the child incorrectly offered the objects from the referred box, E1 acted surprised and said, "Humph.

Strange. No, that is not what I meant.” He then looked in the other box and said, “Ah. That is what I meant. How did it get here?” and they played with the objects from the non-referred box.

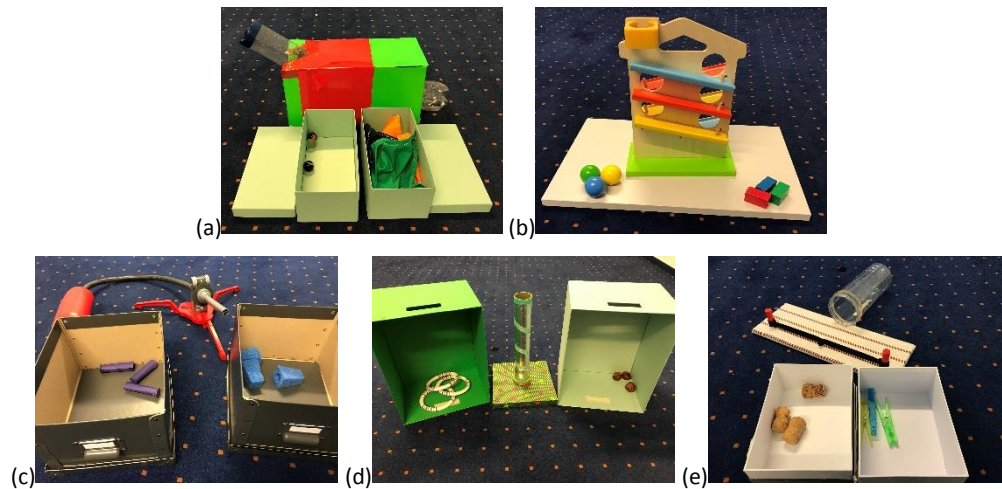


Figure 9. Objects and apparatuses used in the familiarization trials (a, b) and test trials (c – e) of the pragmatically modified sefo task.

Results

All statistical tests were performed in IBM SPSS Statistics Version 23. Alpha was set at 0.05. All presented p-values are two-tailed if not mentioned otherwise.

Direct replication

Direct replication at 17 months of age

To provide a direct replication of the analyses of Southgate et al. 2010 (single trial design), we first present analyses of the first test trial for the 17-month-olds. To ensure that the modifications of the warm-up phase had no influence on test trial performance, we compared the choice of box between those participants who met the original criterion (FB: $n = 12$, TB: $n = 16$) and those who passed with the more lenient criterion (FB: $n = 12$, TB: $n = 8$). There was no difference in the choice of box between the groups in each condition (Fisher’s exact tests, FB condition: $p = 1$, TB condition: $p = .667$), and within each group no difference between FB or TB

condition (Fisher's exact tests, original criterion: $p = 1$, modified criterion: $p = 1$). Thus, we collapsed all participants for further analyses.

First trial

The 17-month-olds performed at chance in the FB condition of the sefo task, only 9 of 24 (38%) correctly chose the non-referred box (binomial test, $p = .307$). In the TB condition, 16 of 24 participants (67%) correctly chose the referred box, which was also not different from chance (binomial test, $p = .152$). Figure 10 shows performance in both conditions. In contrast to the original study, there was no difference in the number of infants who chose the referred box between the TB and the FB condition (Fisher's exact test, $p = 1$). Combining the two conditions shows that overall infants tended to choose the referred box (binomial test, $p = .059$). There were no effects of sex or lab (Fisher's exact tests, all $ps \geq .667$).

Repeated trials

39 participants took all three trials, 7 took two trials (5 in FB, 2 in TB), and 2 took only one trial (both in FB). In the FB condition, in the second trial, 64% chose the non-referred box, and in the third trial, 24%. 63% and 48% chose the referred box in the TB condition, respectively. Within each condition (TB or FB), there was no difference in performance between the three test trials (McNemar tests, all $ps \geq .180$), except for a trend in the FB condition that performance reduced from the second to the third trial (McNemar test, $p = .065$). Within each test trial, there was no significant difference in performance between FB and TB condition (Fisher's exact tests, all $ps \geq .100$).

Sefo task and comparison to standard task at 3 years of age

In the FB condition of the sefo task, 8 of the 24 3-year-olds (33%) correctly chose the non-referred box, which was not different from chance (binomial test, $p = .152$). In the TB condition of the sefo task, 22 of 24 (92%) correctly chose the referred box, which was significantly different from chance (binomial test, $p < .001$). The 3-year-olds tended to choose differently in the two conditions of the sefo task (Fisher's exact test, $p = .072$). A comparison of performances in the sefo task between 17-month-olds (first trial) and 3-year-olds showed no difference in the FB condition (Fisher's exact test, $p = 1$). In the TB condition, 3-year-olds tended to perform better compared to 17-month-olds (Fisher's exact test, $p = .072$).

In the FB condition of the standard task, 13 of the 24 children (54%) correctly chose the empty box, which was not different from chance (binomial test, $p = .839$). In the TB condition of the standard task, 11 of 24 (46%) correctly chose the full box, which was not different from chance (binomial test, $p = .839$). The standard task conditions were not different from each other (Fisher's exact test, $p = 1$). A comparison of performances between the sefo task and the standard task showed no significant difference and no correlation in the FB condition (McNemar test, $p = .267$; $\phi(24) = -.059$, $p = .772$) but a significant difference and no correlation in the TB condition (McNemar test, $p = .003$; $\phi(24) = -.025$, $p = .902$). Figure 10 shows performance of the 3-year-olds in each condition of both tasks. There were no effects of sex, lab, or task order (Fisher's exact tests, all $ps \geq .729$), and no correlation with age (sefo: $r(48) = -.012$, $p = .933$; standard: $r(48) = -.024$, $p = .873$) for the choice of box in each task.

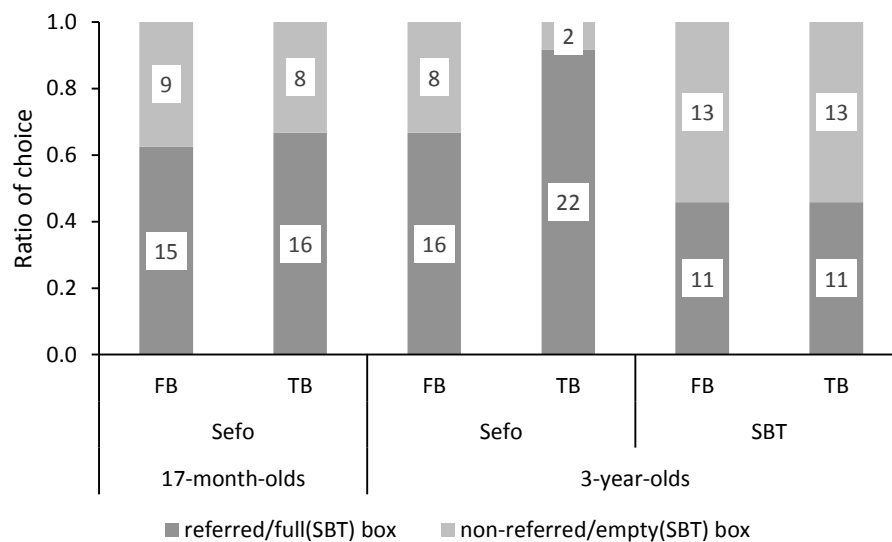


Figure 10. Proportion of participants (17-month-olds (first trial performance depicted) and 3-year-olds) who chose each box in the false belief (FB) and the true belief (TB) condition in the direct replication of the sefo task and standard belief task (SBT). Numbers in bars show number of participants.

Pragmatically modified task

First trial

All infants correctly brought the indicated objects in the two familiarization trials. Since the FB feedback condition and FB no feedback condition only differed in procedure after

children’s approach in the first trial, and since there was no significant difference between performance in the two conditions (Fisher’s exact test, $p = .640$), we collapsed them for the first trial analyses. In the first trial, in the FB condition, 6 of the 24 children (25%) correctly chose the non-referred box, which was significantly different from chance (binomial test, $p = .023$). In the TB condition, 10 of the 12 children (83%) correctly chose the referred box, which was significantly different from chance (binomial test, $p = .039$). There was no difference between the FB and the TB condition in children’s choice of the referred box (Fisher’s exact test, $p = .691$). There was no effect of sex in each condition (Fisher’s exact tests, all $ps \geq .470$).

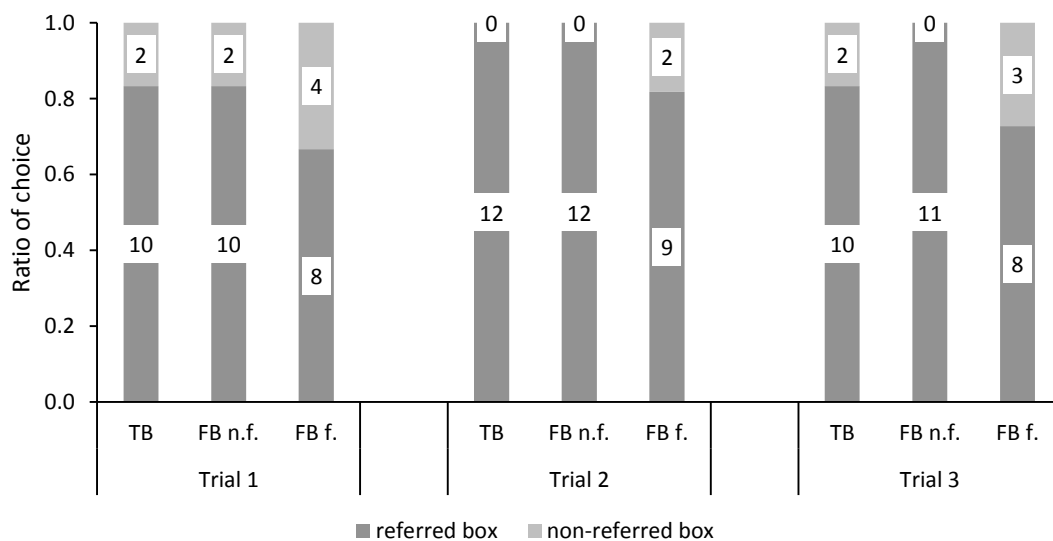


Figure 11. Proportion of participants who chose each box in the three test trials for the true belief (TB), false belief no feedback (FB n.f.) and false belief feedback (FB f.) condition in the pragmatically modified sefo task. Numbers in bars show number of participants.

Repeated trials

34 participants took all three trials, one took two trials (FB no feedback), and one took only one trial (FB feedback). Figure 11 depicts performances in all test trials for each condition. Within each condition (TB, FB feedback, FB no feedback), there was no difference in performance between the three test trials (McNemar tests, all $ps = 1$). Within each test trial, there was no difference in performance between the three conditions (Fisher’s exact tests, all $ps \geq .214$).

Discussion

Our study failed to reproduce findings that infants at 17 months of age ascribe beliefs to others in an interactive FB task. In the original study by Southgate et al. (2010), 17-month-olds correctly interpreted the communicative reference of an experimenter and chose the non-referred box above chance level in the FB condition, while choosing the referred box in the TB condition. Contrary to the original study, we found that the majority of participants in the sefo task chose the box that was indicated by the experimenter irrespective of whether she held a FB or a TB. Our findings are in line with other studies that found similar results and failed to replicate the sefo task at older ages (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2017).

However, the current study found that infants' performance in both test conditions of the sefo task was not different from chance level. This may be due to a power problem, since we found a statistical trend for choosing the referred box when combining the FB and TB condition, increasing the sample size. Findings from our modified version of the sefo task suggest that infants show a clear above chance preference for the referred box in a more pragmatic context that gives specificity and reason to the request of the experimenter. As in our direct replication, infants in that modified task did not consider the belief of the experimenter when making a response. This makes it likely that some infants in our direct replication of the sefo task chose the other box due to uncertainty about the experimenter's goal and triviality of the task. Even across multiple trials with helpful feedback, 17-month-olds in the direct replication and 24-month-olds in the modified task version were not able to pass the sefo task and to overcome their bias to go for the referred box. Thus, our findings strongly question the reliability of the sefo task as a measure of early FB understanding, and raise the possibility that the original findings were false positive.

One limitation of our findings might be that the 17-month-olds showed a weak warm-up performance and were generally only moderately cooperative during that phase (e.g., bringing only one of the objects). We found this behavior equally in the two labs. It is at least a possibility that a low level of cooperativeness or understanding of the general procedure during the warm-up negatively affected test trial performance in our study. However, since those infants who passed the warm-up as in the original study and those who passed with the more lenient criterion performed similarly in the test trials with no statistical difference, it is very unlikely that the warm-up performance could explain why we failed to find the original effects.

Also at 3 years of age, we found chance performance in the FB condition with only about a third of children passing. The sefo task replication study by Grosse Wiesmann et al. (2017) found a similar amount of passers at age 3 although their task had limitations, such as that the second experimenter was in the room during the test trials, which may have decreased the performance enhancing deceptive mode. In our study, however, we followed the procedure of the original study and the second experimenter hid behind curtains. Since both studies found similar results, it is unlikely that the presence of the second experimenter in the study by Grosse Wiesmann et al. influenced performance. The fact that 3-year-olds' performance in our study was not different from performance of the 17-month-olds in our study further reject the notion that the sefo task might be more suitable for testing infants than toddlers.

Contrary to Grosse Wiesmann et al. (2017) who found a trend for a weak correlation, performances between standard and sefo task were not correlated in our study at age 3. This suggests that both tasks make different demands. Although not statistically significant, our 3-year-olds performed weaker in the FB condition of the sefo task compared to the standard FB task (33% passers compared to 54%, respectively). The pattern in our standard FB task is in line with meta-analytic findings (Wellman et al., 2001), suggesting a passing rate of about 50% at 3.5 years (which is the mean age in our sample). Recent studies found that chance performance in a SFB task at age 3 does not indicate merely that children choose randomly, but that those passing the standard task may use their ToM competence (Dörrenberg, Wenzel, Proft, Rakoczy, & Liszkowski, 2019; Lohmann, Carpenter, & Call, 2005). This makes it even more puzzling why these young standard FB passers did not perform equally or better in the sefo task (but even worse), because non-verbal FB tasks are generally considered to be easier and require less cognitive effort due to reduced inhibitory or language demands (e.g., Scott & Baillargeon, 2017).

Interestingly though, two studies found that performance in the sefo task did not correlate with performance in other implicit FB tasks, such as anticipatory looking tasks (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2017). A reason for that might be that the sefo task makes higher cognitive and linguistic demands compared to other implicit tasks. For instance, the request of the experimenter elicits a choice between the two boxes which both contain an interesting toy. An unambiguous understanding of that request requires not only to interpret the pointing gesture but also the verbal prompt. Thus, passing the sefo task may require advanced inhibition, decision making and language skills. This high load makes it unlikely that children utilize an implicit mindreading system for solving the task, as two-systems accounts

would suggest (Apperly & Butterfill, 2009; Low et al., 2016). Regarding demands, the sefo task appears to be more comparable to standard explicit ToM tasks.

Then why are the two kinds of FB tasks not correlated and why did the 3-year-olds perform weaker in the sefo task? One explanation might be that the sefo task makes even higher demands than the standard task. That is, instead of asking an explicit test question as in the standard task, the experimenter in the sefo task points at a box when requesting an object. Several studies have found that 3- to 4-year-olds exhibit a bias to search in pointed-to locations even when the pointer is unreliable, deceptive or ignorant about the object location (e.g., Couillard & Woodward, 1999; Palmquist, Burns, & Jaswal, 2012; Palmquist, Kondrad, & Norris, 2018; Palmquist & Jaswal, 2012; Povinelli & De Blois, 1992). In a study by Palmquist and Jaswal (2012), for instance, children saw a video in which one actor hid an object under one of two cups, while another actor covered her eyes and did not see the hiding location. The hider used a barrier so that also the child was ignorant. After the hiding, children were asked who of the two actors knew where the object was. When both actors sat with their hands in their lap during the test question or each grasped the top of a different cup, children significantly selected the actor that hid the object. When both actors pointed at a different cup, though, children selected at chance level and did no longer discriminate between the knowledgeable and the ignorant actor. This indicates that pointing leads children to attribute knowledge also to obviously ignorant agents.

Accordingly, even if children would understand the false belief of the agent in the sefo task, they are still likely to fail due to problems in inhibiting their bias to search in pointed-to locations. Importantly, already young infants at about 12 or 14 months of age can infer the intention of a pointer to have them look in the indicated location (Behne, Carpenter, & Tomasello, 2005; Behne, Liskowski, Carpenter, & Tomasello, 2012). This makes it likely that the bias to search in pointed-to locations also applies to infants at the age of the original sample. Thus, the necessity to choose against a pointing gesture makes the sefo task an unsuited measure of FB understanding in infants and toddlers.

In a seeing-blindfold paradigm based on the sefo task, Király et al. (2018) found that 3-year-olds chose the non-referred box significantly more often than expected by chance in the FB condition (using opaque sunglasses), which was significantly different from a TB condition (using transparent sunglasses). Their findings contrast the findings of the current study and of Grosse Wiesmann et al. (2017), who found that the majority of 3-year-olds chose the referred box in the FB condition of the standard sefo task. However, in contrast to Király et al. (“Do you remember

what I put in here?”), the other replication studies used a phrasing of the experimenter’s request that did not refer to the child’s memory of where the experimenter put the object before the swapping (e.g., “Do you know what’s in here?”). Thus, it is likely that participants in the study by Király et al. made use of a literal interpretation of the request (i.e., bringing the object that they remembered E1 putting in the referred box before it was swapped by E2) and not of FB understanding. One argument against this interpretation might be that in the TB condition participants chose the referred box, although the verbal prompt should equally guide them to the non-referred box. However, maybe the fact that E1 was virtually part of the object swapping in the TB condition (she was attentive and did not protest against it) made them to interpret the prompt differently (in terms of “Do you remember what was put here?” or “Do you remember what we put here?”). Ultimately, it remains an empirical question.

We found that 3-year-olds performed better in the TB condition of the sefo task compared to infants and, thus, tended to choose differently between TB and FB condition. However, the condition difference was only a statistical trend and the majority of 3-year-olds chose the referred box in the FB condition. There are certainly alternative explanations for a condition difference in the sefo task. For instance, the conditions differ regarding trickery, i.e. E1 was tricked in the FB but not in the TB condition. Thus, children in the FB condition could expect a person that was tricked to look in the wrong location and draw her attention to the correct location, just like they do in hide-and-seek games (Perner, 2014). However, some of the 3-year-old participants likely had a ToM competence (e.g., Dörrenberg et al., 2019). The difference between the TB and the FB condition may indicate that those participants passing the sefo task were able to overcome their bias to search in point-to locations, maybe due to better inhibition skills, and to reveal their ToM competence in the sefo task. Our finding of chance performance of 3-year-olds in the standard TB task fits other studies that found the same pattern at similar age (Fabricius, Boyer, Weimer, & Carroll, 2010; Oktay-Gür & Rakoczy, 2017; Perner, Huemer, & Leahy, 2015). A recent study (Oktay-Gür & Rakoczy, 2017) found that failures in standard TB tasks are due to triviality of the test situation in which everyone (experimenter, child, agent) knows about the real state of affairs. Once the standard TB tasks were made less trivial, children passed.

In line with other replication studies, the current study, although conducted as closely as possible to the methods of the original study, failed to reproduce findings with the sefo task of early FB understanding between one and three years of age. Children of all age groups exhibited a strong preference to choose the box indicated by the experimenter. Even across repeated trials

with helpful feedback and in a more pragmatic version of the sefo task, children were not able to choose against the reference of the experimenter. Additionally, standard explicit and sefo task were not correlated at age 3, suggesting different task demands. Passing the sefo task requires children to overcome their bias to search in pointed-to locations, which produces high inhibitory demands. Thus, the sefo task appears to be an unsuited measure of FB understanding and should be handled with caution.

Acknowledgement

This study was supported by the German Research Foundation (LI 1989/3-1, RA 2155/4-1, Project: FOR 2253). We are grateful for the participation of all the parents and children. Further, we want to thank Annika Braun, Rocío Fernandez, Anna Fink, Marc Heuser, Joana Lonquich, Laura Meier, Senta von Münchow, Rieke Oesterreich and Julia Ruge for help with data collection and recruiting participants, as well as Marlen Kaufmann, Konstanze Schirmer and Jessica Schröter for lab coordination.

Study 3: Reliability and generalizability of an acted-out false belief task in 3-year-olds

This study was published in the Journal *Infant Behavior and Development* (Dörrenberg et al., 2019).

Abstract

The current study tested the reliability and generalizability of a narrative act-out false belief task held to reveal Theory of Mind (ToM) competence at 3 years of age, before children pass verbal standard false belief tasks (the “Duplo task”; Rubio-Fernández & Geurts, 2013, *Psychological Science*). We conducted the task across two labs with methodologically improved matched control conditions. Further, we administered an analogue intentionality version to assess the scope of ToM competence in the Duplo task. 72 3-year-olds participated in a Duplo change-of-location task, a Duplo intentionality task, and half of them in a matched verbal standard change-of-location task, receiving either false belief or matched true belief scenarios. Children performed at chance in the false belief Duplo location change and intentionality tasks as well as in the standard false belief task. There were no differences to the standard task, and performance correlated across all three false belief tasks, revealing a rather unified competence and no task advantage. In the true belief conditions of both Duplo tasks, children performed at ceiling and significantly different from the false belief conditions, while they were at chance in the true belief condition of the standard task. The latter indicates that a pragmatic advantage of the Duplo task compared to the standard task holds only for the true belief scenarios. Our study shows that the Duplo task measures the same ToM competence as the standard task and rejects a notion of earlier false belief understanding on the group level in 3-year-old children.

Introduction

Theory of Mind (ToM), the ability to attribute mental states to others, is typically tested with false belief (FB) tasks that require the ascription of others’ subjective representations of reality that can be false (Wimmer & Perner, 1983). In change-of-location tasks, for example, children see a protagonist put an object in one of two boxes. In the protagonist’s absence, the object is

then transferred to another box and children are asked to predict where she will look for it. These standard verbal tasks are mastered from age four while young 3-year-olds typically fail (by predicting that the protagonist will look for her object where it really is) and group performance at three-and-a-half years is typically at chance (Wellman et al., 2001). Since these tasks require advanced pragmatic understanding of language and test questions, several studies have lowered these demands and found enhanced performance at slightly younger ages (e.g., Mitchell & Laco  e, 1991; Psouni et al., 2018; Rhodes & Brandone, 2014; Sullivan & Winner, 1993; for a meta-analytic finding, see Wellman et al., 2001). A recent study employed the “Duplo task” (Rubio-Fern  ndez & Geurts, 2013), and reduced demands by minimizing disruptions during the perspective tracking process. Using a narrative version of the change-of-location FB paradigm, children were prompted to act out the protagonist’s action, instead of answering to an experimenter’s explicit test question about the protagonist’s belief. Furthermore, the protagonist remained visible throughout the narrative to facilitate keeping track of her perspective. With these variations, 3-year-olds performed in the Duplo task above chance while still failing a FB task with an explicit test question. The current study sought to test (i) the robustness of the finding through a multi-lab replication approach and by implementing methodological improvements, and (ii) the scope and unity of early ToM competence by administering an intensionality version of the Duplo task, and testing for correlated performance across tasks (see Rakoczy et al., 2015).

In the original Duplo task, Rubio-Fern  ndez and Geurts (2013) introduced two main modifications to the standard false belief (SFB) task to facilitate children’s perspective tracking. First, children could undisturbedly keep track of the agent’s knowledge access during the whole story: The protagonist did not leave the scene at all, but turned her back towards the scenery, so that she was unaware about the object’s transfer, but still visible to the child throughout this procedure. Additionally, two prompts about her knowledge access (e.g. “She hasn’t seen what I did, did she?”) should help children to keep track of the protagonist’s perspective. Second, they used a narrative story structure in which children were not asked explicitly where the protagonist would search for the object, but they were involved actively and encouraged to act out the story (“What happens next? You can take the girl yourself if you want... What is she going to do now?”). While 80% of children passed the novel Duplo task by acting out the belief-congruent ending (moved the Duplo girl to the container where she previously left her banana), the same proportion of children failed a standard unexpected-content task (Hogrefe et al., 1986). In two

follow-up experiments, Rubio-Fernández and Geurts (2013) showed that these two modifications crucially led to the increased performance in 3-year-old children. If the protagonist disappeared from the scenery during the location change, or if children were asked an explicit test question that mentioned the desired object instead of being given the actively engaging prompts, children's performance decreased to below chance. In another study (Rubio-Fernández & Geurts, 2016), the same authors investigated the impact of two further task modifications. A modification of the test question ("Where will Lola go now?" instead of "What happens next?"), which may highlight the binary choice between the two locations and thus increase attention towards the box containing the object, had no effect on children's increased performance in the Duplo FB task. A stressed salience of the target object, however, decreased children's performance: When mentioning the object, either in a control question after the transfer ("Where are the *bananas* now?"), or right before children were prompted to take the lead ("Now Lola is very hungry and wants a *banana*."), 3-year-old children failed to solve the Duplo FB task. Given the set of modifications implemented in the Duplo task and their implications on children's performance to pass or fail FB tasks, it becomes clear that 3-year-olds' perspective tracking skills are still fragile and can be both enhanced and disrupted quite easily by subtle but crucial factors. Findings of 3-year-old children passing simplified FB tasks have been taken to support early competence accounts and to suggest that extraneous task demands mask false belief understanding in younger children (Carruthers, 2013; Leslie, Friedman, & German, 2004; Scott & Baillargeon, 2017). While the interpretation has far-reaching implications for the origins and nature of ToM competence, it is important to first assess the validity and robustness of the empirical findings by replicating the original studies. Concerning the Duplo task, to date, there are two published studies that conducted conceptual replications with several (some substantial) modifications of the original protocol, one of which comes from an independent lab (Białecka-Pikul, Kosno, Białek, & Szpak, 2019; Rubio-Fernández & Geurts, 2016). These studies found enhancing effects and above chance performance in their Duplo task versions compared to standard verbal FB tasks. The task by Białecka-Pikul et al. (2019) retained only some minor elements of the Duplo task (e.g. the knowledge access prompts), but introduced a more interactive "we-mode" (child and experimenter jointly tricked the agent). In contrast to the original version, children were asked an explicit test question and the target object was mentioned. These are both factors which should have decreased performance on the task according to Rubio-Fernández and Geurts (2013, 2016), thus questioning the validity of the

original task manipulation. A more direct replication of the Duplo task (Kammermeier & Paulus, 2018) revealed a facilitating effect of the Duplo task compared to a verbal FB task, but in contrast to the original findings, 3-year-olds performed only at chance in the Duplo replication task. This finding questions the reliability of the original above chance finding in the Duplo task and calls for clarification.

Unfortunately, there were also some methodological limitations to the original Duplo task itself. First, the true belief (TB) control condition was not matched to the FB condition. In contrast to the FB condition, in the TB condition the Duplo girl moved the banana herself, and was never distracted from the events, which makes the TB demands easier and prone to a leaner interpretation like a simple agent-object association (e.g., Perner & Ruffman, 2005). Second, for the benchmark comparison to the explicit verbal ToM competence, the authors used an unexpected-content task which does not match the change-of-location structure of the Duplo task (this was the same in Białecka-Pikul et al., 2019). While key manipulations of the Duplo task concern perspective tracking and mentioning of the target object, unexpected-content tasks actually do not involve comparable perspective tracking (there is no agent), and do not bias children to the wrong container via test question (the target object is not mentioned in the test question, “What will XY think is in the box?”). This makes the task less ideal for the investigation of these manipulations. Further, from a conceptual point of view, although performance in change-of-location and unexpected-content tasks correlate (Gopnik & Astington, 1988), the cognitive demands differ between tasks. For example, unexpected content tasks may be more difficult than location-tasks, because they are even more language-dependent, provide less supportive story context, and require transferring one’s own FB to another person. From an empirical point of view, there is indeed evidence that unexpected-content tasks are more difficult than change-of-location tasks (Gopnik & Astington, 1988; Holmes, Black, & Miller, 1996). Similarly, the direct replication study of the Duplo task (Kammermeier & Paulus, 2018) did not use an adequately matched change-of-location FB task for performance comparison. In their task, children were only told that the protagonist thought his mittens were in a closet even though they were in his backpack, but they did not see the actual location change (Wellman & Liu, 2004). Meta-analytic findings have shown that at three-and-a-half years (which is the mean age group of participants in the original Duplo task study and in all replication studies), children perform at chance (with about 50% passing rate) in verbal FB tasks (Wellman et al., 2001). Compared to this, children performed rather poorly in the verbal location FB task in the

replication study by Kammermeier and Paulus (at age 3: below chance, at age 4: at chance), rendering it possible that the facilitating effect of the Duplo task was rather an artefact of the poor performance on the non-matched SFB task.

Taken together, methodological differences make it thus difficult to interpret the performance in the Duplo FB task relative the TB condition and the employed verbal FB tasks. It remains to be tested, how the findings from the Duplo task compare to more closely matched control conditions. A recent study on FB understanding in a word learning context addressed the issue of miss-matched conditions (Papafragou et al., 2017). When the communicative and the non-communicative FB conditions were matched exactly, earlier findings of an advantage of word learning tasks (Carpenter et al., 2002; Happe & Loth, 2002) could not be reproduced. This highlights the importance of comparable controls.

Another important question is what the Duplo task actually measures. After the age of four, children master a variety of explicit ToM tasks in converging fashion. They not only solve classical change-of-location FB paradigms, but also solve unexpected-content, appearance-reality and intensionality tasks (measuring the understanding that someone has a false belief about different aspects of an object), and there is inter-task coherence (Gopnik & Astington, 1988; Perner & Roessler, 2012; Rakoczy et al., 2015). Thus, 4-year-olds seem to have a fully-fledged, unified and flexible ToM competence. This convergence does not apply to implicit ToM tasks that use non-verbal measures. Here, different tasks are not equally solved, but show a dissociation, i.e. children master implicit change-of-location tasks, but fail implicit intensionality tasks (Fizke et al., 2017; Low & Watts, 2013; Oktay-Gür et al., 2018). Implicit ToM tasks may therefore tap into an earlier developing and efficient mindreading system, which might be the developmental basis for later flexible ToM, as proposed by two-systems accounts (Apperly & Butterfill, 2009; Low et al., 2016). This early ToM system may, according to two-systems views, only be capable of tracking belief-like states or simpler perceptual registrations (what someone saw or did not see), and exhibit specific signature limits, such as the ascription of false beliefs about aspects of objects. For mastering the Duplo task, however, it is unclear whether 3-year-olds utilize a fully-fledged and unified ToM competence that is camouflaged in classical tasks, or whether their ToM competence is immature and exhibits limitations.

Against this background, the rationale of the present study was to test the robustness and reliability of the performance enhancing factors implemented in the Duplo task, and the scope of the underlying ToM capacity. In order to check for robustness and reliability, we tested

3-year-olds in a change-of-location version of the Duplo task and implemented further methodological changes. In contrast to the original study, we conducted a TB condition that closely matched the FB condition and was equated in terms of performance factors. Therefore, both conditions equally required children to track the protagonist's mental state. The only difference between conditions was that in the FB condition the protagonist turned around somewhat later and thus did not witness the crucial event (location change). In addition, we implemented the feature of joint trickery, in order to increase children's involvement throughout the story (Białecka-Pikul et al., 2019; Sullivan & Winner, 1993). Instead of an unexpected-content task as benchmark comparison, we administered a change-of-location task, which matched the features of the Duplo task, but was different in two crucial respects. First, we used an explicit test question as dependent measure (mentioning the desired object), instead of encouraging children to act out the story. Second, rather than turning his back at the scenery during the location change, the protagonist left the scene, as in standard false belief tasks. Following the Duplo task logic, and assuming the task to be robust and reliable, we should find correct above chance performance in the FB condition of the Duplo task, and a differential choice pattern between TB and FB conditions. Further, children should perform differently and less competently in the SFB compared to the Duplo task FB condition.

To find out whether there are limits in 3-year-olds' Duplo task performance, or whether it measures a fully-fledged ToM competence, we designed an intensionality version of the Duplo task that was identical to the change-of-location version with one exception: Rather than failing to witness the location change in the FB condition, the protagonist failed to witness how the experimenter revealed that the object (a pen) also had a second identity (is also a rattle). The protagonist then saw how the object, under this second identity (as a rattle), was transferred to the other container (Rakoczy et al., 2015). If there is unity, children should solve the intensionality version on the same level as the change-of-location version, and there should be convergence and correlation between them. We also tested for correlations between the Duplo task versions and the SFB task, which has not been done yet, in order to clarify if they measure the same ToM capacity.

Methods

Participants

The final sample of the study included 72 3-year-old children (36-47 months, $M = 41.6$, $SD = 3.1$; 31 boys) from mixed socioeconomic background. The data collection was conducted in two different labs (first in Hamburg ($n = 32$), then in Göttingen ($n = 40$)) each by a female experimenter. In Hamburg, children were recruited and tested in seven different nurseries with written consent of their parents. In Göttingen, data collection was conducted either in children's nursery or in the lab and were recruited from databanks of children whose parents had previously agreed to participate in child studies. Four children were excluded because they did not cooperate, and testing sessions had to be interrupted.

We conducted statistical power analyses for sample size estimation (using G*Power 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007), aiming at $\alpha = .05$ and power = 0.95, based on data from the original Duplo task study ($N = 25$; Rubio-Fernández & Geurts, 2013). The effect size in the original study for the above chance performance in the Duplo task was 0.3. The projected sample size needed with this effect size is approximately $N = 35$. For the performance improvement compared to the verbal FB task, the effect size in the original study was 1.15. The projected sample size needed with this effect size is approximately $N = 15$.

Design and Procedure

We presented each child with two versions of the Duplo task: A change-of-location task and an analogue intensionality task (within-subject; modeled after Rakoczy et al., 2015). Children received either TB or FB scenarios (between-subject). The tasks were presented in counterbalanced order. The Göttinger children additionally received a standard change-of-location task in the same condition (TB/FB) as the Duplo task versions (Wimmer & Perner, 1983). Here, the order of task type (standard and Duplo versions) was also counterbalanced. Children were tested individually in a quiet room. Each session started with a warm-up phase, in which the child and the experimenter (E) played with a cuddly toy and some objects. Already in this phase, children were encouraged to slip into the role of the puppet and act on her behalf.

Duplo tasks

The procedure of the Duplo change-of-location task was adopted from Rubio-Fernández and Geurts (2013) and analogously adapted for the intensionality task (after Rakoczy et al., 2015). Since the original intensionality task included explicit trickery together with the child (see Rakoczy et al. 2015, Appendix A), a feature that increased performance in other FB tasks (e.g., Białecka-Pikul et al., 2019; Sullivan & Winner, 1993), we implemented this feature in all of our tasks. Note that also in the original Duplo task the experimenter acted as if she would play a trick on the protagonist. During creation and piloting of the experimental protocol, we were in regular contact with the first author, P. Rubio-Fernández.

Duplo change-of-location task

E sat next to the child at a table and acted a puppet story. Before she started, E noted that at the end of the story she would need some help from the child. First, two boxes (representing fridges; materials are depicted in Figure 12) and two objects (an apple and a banana) were shown to the child. The second object, the apple, was not used in the original study. However, in our design, it served to equate the demands of the two task versions, since the intensionality version included an object with two identities (see e.g., Fiske et al., 2017). Next, the protagonist, a cuddly toy ape (either the chimp Freddy, or the orangutan Klaus, counterbalanced across trials), was introduced and expressed his obsession with bananas (E mimicked the ape's voice). After he joyfully discovered the banana next to the fridges, he declared his plan to eat it right after his return, in order to stress his intention for the end of the story, and placed it together with the apple in one of the two fridges (side counterbalanced). He then walked across the table, passed the child and sat with his back turned to the fridges beside the scenery. Different to the original study, the ape put on little headphones when he turned away, to make sure that he could not hear what would happen. This was especially important for the novel intensionality version, which included a toy that made noise (rattling). In both conditions, when the ape was beside the scenery, E proposed to play a trick on the protagonist together in a sneaky manner ("Do we want to play a trick on Freddy/Klaus?"). In the FB condition, E then asked whether the ape was able to see or hear what the child and E were doing, in order to draw the participant's attention to the ape's perceptual state. In the TB condition, on the contrary, the ape turned back too early, and E asked about his perceptual access while he was already sitting in front of the two fridges. E reacted always in a confirmative (for correct answers) or a corrective (for incorrect answers) manner, raising the awareness of the ape's attention to

the scene, which was either absent (FB) or present (TB). E then moved the object from one box to the other, and said, "Look, I put the banana in this fridge." After the transfer, E asked whether the monkey witnessed the object's location change ("Did Freddy/Klaus see that we put the banana in the other fridge?"). Note that, unfortunately, our phrasing of this prompt deviated from that in the original study ("She hasn't seen what I did, has she?"). E reacted in either a confirmative or a corrective manner to the child's answer, pointing to the knowledge (TB) or the ignorance (FB) about the transfer (e.g., "No, he hasn't seen what happened."). In contrast to the original procedure, the story was re-acted from the beginning if children answered incorrectly, to make sure they followed the narrative (this modification was conducted in the Göttinger sample only, as an improvement due to a few children's failure to answer this question correctly in the Hamburger sample). While the protagonist was already present during the transfer in the TB condition, the ape returned at this point of the story in the FB condition and was positioned in front of the two fridges. In order to actively engage the participant in taking over the ape's action, E uttered the question, "Humph, what happens next? Can you help me? Can you now take over Freddy/Klaus and continue playing the story?" If children did not cooperate in the first place, E prompted further, "What is Freddy/Klaus going to do?" If the prompt was also insufficient, she asked, "Will he approach one of the fridges?"

Duplo intensionality task

The intensionality task followed the very same procedure as the Duplo change-of-location task and was modeled after Rakoczy et al. (2015). The protagonist was introduced and stated his obsession with painting. Joyfully discovering a new pen he never saw before, he placed it into one of two toy boxes and stated his intention to paint when he comes back, at the same time showing that the box was otherwise empty. He then walked across the table, passed the child, put on the headphones, and sat with his back turned to the toy boxes beside the scenery. E then asked the child if they would want to play a trick on the protagonist and stated that she would share a secret with the child (acting in sneaky manner), namely that the pen had a second non-obvious identity as a rattle when shaken ("Look, the pen is also a rattle!"). In the TB condition, the ape turned around too early, so the secret was also shared with him. Thus, in contrast to the FB condition, he knew about the object's second identity. The child's awareness about the protagonist's states of perception and knowledge were raised with questions similar as in the Duplo change-of-location task described above. In the presence of the protagonist, E transferred the object into the other toy box under the second identity: she covered the object

with her hand (so that it was not visible), shook it during the transfer (so that its rattle-identity became perceivable), and said, “Look, I put the rattle in this box.” After the transfer, children were engaged in acting out the story as described in the Duplo change-of-location version.

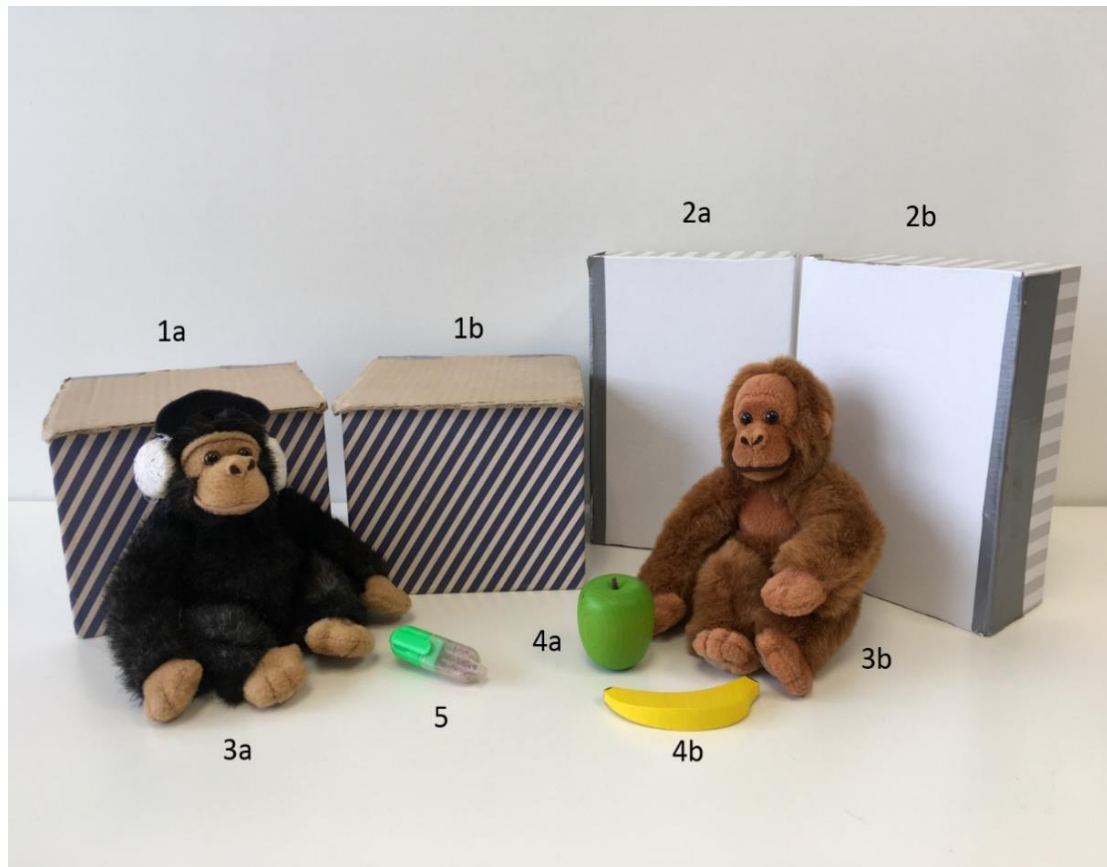


Figure 12. Materials used in Duplo tasks: (1a, b) toy boxes and (5) the rattling pen used in the Duplo intentionality task, (2a, b) fridges and (4a, b) objects used in the Duplo change-of-location task, as well as both protagonists, (3a) the chimp Freddy wearing the headphones and (3b) the orangutan Klaus.

Standard change-of-location task

The standard change-of-location task was adapted to the narrative puppet play of the Duplo task, but crucial factors responsible for disrupted perspective tracking (e.g. the visual absence of the protagonist during the swap, or mentioning the target object in an explicit question; Rubio-Fernández & Geurts, 2013, 2016) were included. A cuddly toy lynx (Luchsi) put his car into one of two containers and left the scene, so he was not visible for the child. E then

proposed to play a trick together on the protagonist in the same way as in the Duplo tasks. In his absence (FB), or after his return (TB), E swapped the car to the other container in a sneaky manner. Children got the same prompts of the protagonist's perceptual and knowledge states as in the Duplo task. However, instead of actively engaging the children to act out the end of the story, an explicit test question mentioning the target object was asked ("Where will Luchsi look for his car first?").

Results

The Duplo tasks

In the location change version, six of the 72 participating children gave ambiguous or no answer and were thus excluded from the main analysis. Figure 13 depicts the percentage of children's answers as a function of task and condition. In the TB condition, 94% of the children acted out the belief-congruent ending of the story and placed the protagonist in front of the box containing the target object (binomial test, test value = 0.5, 29 out of 31 correct, $p < .001$, two-tailed; see Table 2). In the FB condition, 51% chose the belief-congruent box that did not contain the target object (further referred to as empty box) for the story ending (binomial test, test value = 0.5, 18 out of 35 correct, $p = 1$, two-tailed). The difference between the conditions in selecting either the full box or the empty box was significant ($\chi^2(1, N = 66) = 15.75, p < .001$, two-tailed).

In the intensionality version, four children had to be excluded due to ambiguous or no answers. In the TB condition, 85% of the children chose the belief-congruent box containing the object (binomial test, test value = 0.5, 29 out of 34 correct, $p < .001$, two-tailed). In the FB condition, 53% of the children chose the belief-congruent and thus empty box (binomial test, test value = 0.5, 18 out of 34 correct, $p = .864$, two-tailed). The difference between the conditions in selecting either the full box or the empty box was significant ($\chi^2(1, N = 68) = 11.1, p = .002$, two-tailed).

A comparison of the performances between the location change and intensionality versions revealed no significant differences for the FB conditions ($p = 1$, McNemar test, two-tailed, $n = 33$) and TB conditions ($p = .375$, McNemar test, two-tailed, $n = 29$). The FB conditions of the location change and the intensionality version correlated significantly with each other

($\phi(33) = .393, p = .024$). The TB conditions did not correlate with each other, due to ceiling effects in each task ($\phi(29) = .236, p = .204$).

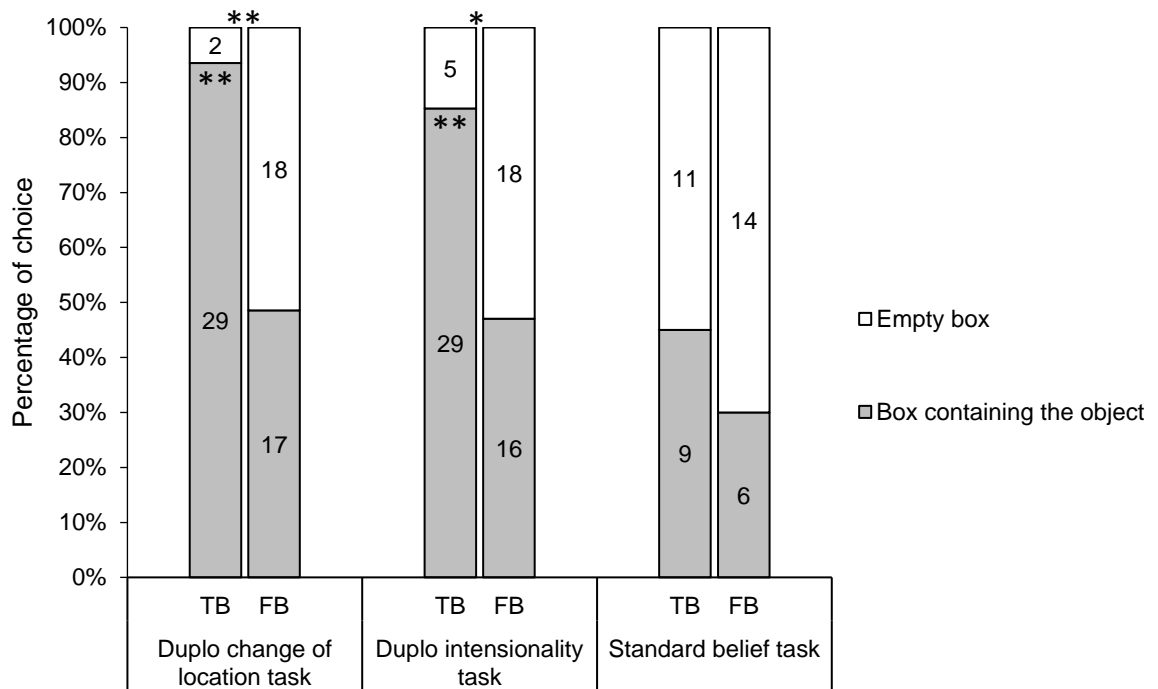


Figure 13. Percentage of chosen box in all three tasks for both true belief (TB) and false belief (FB) conditions. Numbers in bars show number of children. * $p < .01$, ** $p < .001$.

Standard change-of-location task and relations to Duplo tasks

None of the 40 children who received a standard change-of-location task had to be excluded. In the SFB condition, children performed at chance level (binomial test, test value = 0.5, 14 out of 20 correct, $p = .115$, two-tailed). A comparison between the Duplo location change FB and the verbal SFB yielded no differences ($p = .125$, McNemar test, two-tailed, $n = 20$). Further, the Duplo location change FB and the verbal SFB tasks correlated significantly with each other ($\phi(20) = .655, p = .003$). Similarly, a comparison between the Duplo intensionality FB and the verbal SFB revealed no significant differences ($p = .625$, McNemar test, two-tailed, $n = 18$) and a significant correlation ($\phi(18) = .553, p = .019$).

In the TB condition of the standard location change task, children performed at chance (binomial test, test value = 0.5, 9 out of 20 correct, $p = .824$, two-tailed) and there were no significant differences between the TB and FB conditions ($\chi^2(1, N = 40) = 2.56, p = .200$, two-

tailed). Children performed significantly better in the TB conditions of both the Duplo location change task and the intensionality task compared to the TB condition of the verbal standard task (respectively: $p = .004$, McNemar test, $n = 18$; $p = .039$, McNemar test, $n = 19$; both two-tailed), and performance in the TB Duplo tasks did not correlate with the standard TB task, again likely due to the ceiling effects in the Duplo task (respectively: $\phi(18) = .217$, $p = .357$; $\phi(19) = .179$, $p = .435$).

Table 2.

Contingency pattern of given answers in the different tasks for false belief (FB) and true belief (TB) conditions.

		Standard change-of-location		Duplo intensionality	
		Empty box	Full box	Empty box	Full box
Duplo change-of-location					
FB	Empty box	10	0	12	5
	Full box	4	6	5	11
TB	Empty box	1	0	1	1
	Full box	9	8	4	23

Comparisons between the two labs

Preliminary analyses revealed no differences between the two labs in sex ($\chi^2(1, N = 72) = 0.01$, $p = 1$, two-tailed), age ($t(70) = -0.81$, $p = .432$, two-tailed) or overall performance per task (Duplo change-of-location: $\chi^2(1, N = 66) = 0.001$, $p = .973$; Duplo intensionality: $\chi^2(1, N = 68) = 0.09$, $p = .762$; both two-tailed). We therefore collapsed the data sets for the main analyses. However, we also found the same pattern of results as reported above in both labs separately, except for the significant correlation between the two Duplo tasks, which was found only in the Hamburger sample. However, with the increased power of both samples the correlation of performance in the two Duplo tasks was affirmed.

Discussion

The aims of the present study were to clarify how robust and reliable earlier Duplo task findings are and whether the Duplo task measures fully-fledged ToM competence in 3-year-olds. The

main findings were the following. First, we could not reproduce the above chance performance in the Duplo change-of-location version. Second, performance in the Duplo task FB versions was not different from the SFB, and the tasks were correlated. Third, in the Duplo TB conditions, children performed competently, and there was a significant difference to the Duplo FB conditions. Fourth, performance in the Duplo intensionality version was on par with the change-of-location version, and both versions were correlated.

Now, how robust and reliable are the Duplo task findings? Earlier findings of an above chance performance in a narrative FB task could not be reliably reproduced in the present study, which is in line with another recent study (Kammermeier & Paulus, 2018). Why did we find different results? One possibility may be that methodological issues played a key role. Although we adopted the procedure of the Duplo task as closely as possible, and the original first author checked and confirmed the accuracy of our implementation, there were some differences. First, we introduced a second object (an apple) in the Duplo change-of-location task, to equate the demands to the intensionality task that included an object with two identities. Since the Duplo task uses an open response design, with no further instructions about the storyline's direction, this second object may have distracted or confused our participants. However, this seems very unlikely because the apple always remained in the first (and correct) fridge in the FB condition, and thus did not lead children to the incorrect one. Further, not a single child ever referred to the apple in the FB condition when choosing a fridge, thus, there was no indication of confusion. Additionally, in the TB condition, most children correctly chose the location containing the banana, indicating that they understood the protagonist's desire to obtain the focal object, thus ruling out that they were confused by the presence of the apple.

Second, we included one additional feature in half of the sample: we repeated the story when the control question was answered incorrectly. This is a common procedure in explicit ToM tasks, ensuring that children understand the story and reducing the amount of drop-outs (e.g., Clements & Perner, 1994). In our case, we only had to repeat the story for six children, an equal number in each condition (TB or FB). Since we implemented this modification in only one lab and found no differences between the two sample groups, this feature cannot explain the different findings.

Third, there are certainly differences in the interaction style of the experimenters, or in the sample compositions. However, the current study is a collaboration of two labs and testing thus took place in two different cities with two different experimenters. Since both labs found

the same results, this seems to be an unlikely explanation. One informal observation was that children often needed a second prompt to act out the story, which might indicate procedural difficulties in obtaining the dependent measure. Unfortunately, we have no information on similar findings from the other Duplo task studies (Kammermeier & Paulus, 2018; Rubio-Fernández & Geurts, 2013, 2016).

Fourth, we used a slightly different phrasing for the last prompt, during the belief-induction phase (“Did Freddy/Klaus see that we put the banana in the other fridge?” instead of “She hasn’t seen what I did, has she?”). This may have stressed the salience of the target object. Rubio-Fernández & Geurts (2016) showed that emphasizing during the test phase that the protagonist wants the object (“Now Lola is very hungry and wants a banana.”), or asking for the current location of the object (“Where is the banana now?”), can disrupt perspective tracking. Note that our prompt during the belief-induction phase rather emphasized the ignorance of the protagonist, not his desire to get the banana or the current banana location. Note also that German word order buries “banana” in the middle of the sentence rather than exposing it prominently at the end (as in English). Ultimately, it remains an empirical question whether the “banana” in our prompt could explain the worse performance. Findings by Białecka-Pikul et al. (in press) show that mentioning the target object in the test phase has no detrimental effect on children’s performance.

A more general limitation of the Duplo task could be the open response design, which allows for less straightforward predictions of children’s ToM competences than explicit tasks. That is, negative findings do not necessarily speak for a lack of belief ascription abilities. In our Duplo intensionality task, for example, a possible alternative ending could also be that the ape wanted to get the rattle he just found out about; or, in the Duplo change-of-location task that he wanted to check the other fridge. However, as in the original study, we strongly stressed the monkeys’ desire to get the banana/pen at the beginning of the narrative. Most importantly, our TB-control conditions show that children clearly knew how to act out the story coherent with the protagonist’s desire.

While our findings of children’s performance at chance level in the Duplo FB task are in line with the direct replication study by Kammermeier and Paulus (2018), the Duplo-study by Białecka-Pikul et al. (2019), which focused on the influence of interaction in FB tasks, found above chance performance in the FB condition in three-and-half-year-old children and an increased performance compared to a standard unexpected-content FB task. During that task, the

experimenter made the child focus on the protagonist's state of knowledge via prompts just like those in the Duplo task. However, in contrast to the original Duplo task, an explicit test question mentioning the target object was asked, and the protagonist was covered by a cloth during the location change rather than being visible beside the scene. Curiously, then, none of the potentially facilitating main factors of the original Duplo task, namely not mentioning the target object in the test phase, and continuous visibility of the protagonist, were implemented by Białecka-Pikul et al. (2019). Instead, that study implemented an emphasis on a "we-mode" in their procedure that putatively led to children's enhanced performance. That is, the experimenter actively engaged and involved the participating child by using a deceptive motive of joint trickery, saying, "Hey, let's surprise the mouse!" However, in the current study we also implemented the interactive motive of joint trickery in all our tasks ("Do we want to play a trick on Freddy/Klaus?"), in addition to the original key manipulations of the Duplo task, and still found no significant above chance performance. It should be noted, however, that mean performance in Białecka-Pikul et al. (2019) was not drastically different from ours (respectively, 57% correct vs. 51% correct) and certainly diverged strongly from the original finding of 80% correct (Rubio-Fernández & Geurts, 2013). However, Białecka-Pikul et al. (2019) used a six times larger *N* of 210 children, rendering a statistical significance perhaps less meaningful.

Our interpretation of the findings is then that there is no robust facilitating effect of the Duplo task after all. In the unexpected-content task that was originally used as SFB control (and also in Białecka-Pikul et al., 2019) and in the location FB task used by Kammermeier and Paulus (2018), children performed below chance, and thus, performance in the Duplo task was significantly better. However, we found no such difference in performance between the Duplo task and a SFB task when the latter was matched accordingly. It is theoretically possible that the matching artificially enhanced performance in our SFB task, because we included the Duplo prompts during the belief-induction phase about the protagonist's epistemic state (e.g., "Did Luchsi see that we put the car in the other box?"). However, given that the prompts also mentioned the object, it is equally possible that they led to disrupted perspective tracking. Our at chance level findings, however, concur well with the meta-analytic findings (Wellman et al., 2001) suggesting at chance level performance at 3.5 years of age. Though beyond the scope of the current paper, we have recently run our current SFB task in a different study without prompts during the belief-induction phase. We found a very similar pattern of performance at chance with no significant differences to the current results. In addition, other findings show that when

adding an explicit question to the Duplo task (Rubio-Fernández & Geurts, 2013, Experiment 2b), or stressing the target object in the Duplo task (Rubio-Fernández & Geurts, 2016, Experiment 2), performance drops dramatically below chance - despite the epistemic prompts. Thus, it is unlikely that the epistemic prompts alone led to an increase in performance on our SFB task.

Our study equated both kinds of FB tasks in terms of structure and performance factors and implemented an analogue standard change-of-location control that only differed from the Duplo task in the most important elements that disrupt perspective tracking (i.e., agent disappeared from scene, explicit test question mentioning the target object). This is reminiscent to findings of Papafragou et al. (2017) who could not replicate advantages of word learning tasks on FB understanding with matched controls. Further, it urges future researchers to adjust their control conditions carefully. The previously used SFB comparisons were cognitively more demanding than the Duplo task (Gopnik & Astington, 1988; Holmes et al., 1996), which led to the assumption of a facilitating effect. Our minimal contrast design, however, suggests that the Duplo task leads to no improvement and has no facilitating effect when compared to a matching SFB. Indeed, our findings reveal not just that the Duplo task has no facilitating effect – in addition, they reveal that performance on the Duplo task and the SFB task are based on a common capacity. We found evidence for converging performance, i.e. both Duplo task versions correlated significantly with the matching SFB task. Thus, the Duplo task may tap the same ToM competence as standard explicit tasks, and explicit tasks may not underestimate children's FB understanding.

The second aim of the present study was to investigate whether we could transfer the facilitating effect of the Duplo change-of-location task to other FB paradigms, and whether 3-year-olds possess a fully-fledged ToM competence or one that exhibits limitations. As described above, we did not find any facilitating effects in the Duplo change-of-location FB version in the first place. Similarly, performance in the Duplo intensionality FB version was at chance. However, performance of both task versions showed convergence and correlation. Thus, the Duplo task seems to measure a generalizable and not merely local phenomenon. In this respect, it behaves like standard explicit tasks that also reflect this unity and convergence (Perner & Roessler, 2012; Rakoczy et al., 2015), rather than implicit tasks that often show disunity and divergence (Fizke et al., 2017; Oktay-Gür et al., 2018). Performance on both Duplo task FB versions was significantly different from the TB controls. This makes it unlikely that children were totally random in their choice. Given the dichotomous data of the task, a possible interpretation of 3-year-old children's

at-chance performance on the SFB task is that half of the children possess the competence (Lohmann et al., 2005). This interpretation is corroborated by our correlational findings which exclude lower level interpretations that children were just guessing or perseverating. Instead, the pattern of performance across all three FB tasks indicates that children either systematically passed or failed in answering in a belief-congruent way, suggesting that half our participants were competent FB passers. Presumably, these children already possessed a robust and unified ToM capacity that is comparable to that of older children. The other half might have failed, as at the age of three, FB understanding is still a fragile competence that is prone to overwhelming demands on pragmatic confusion and cognitive capacities (Helming, Strickland, & Jacob, 2014; Wellman et al., 2001). The findings thus show that performance in the Duplo task is much closer related to standard explicit tasks than previously assumed.

In line with the original study, we found very good performance in our new matching Duplo TB conditions, and we found significant condition differences. Thus, despite methodological differences, the original TB condition did not overestimate children's performance. In contrast, children performed at chance level in the TB condition of the standard task. On first glance, this finding might seem quite surprising. However, it nicely fits with recent studies that focus on children's TB performance, showing that the relation to FB performance seems more complex than previously assumed (Fabricius et al., 2010; Oktay-Gür & Rakoczy, 2017; Perner et al., 2015). Oktay-Gür and Rakoczy (2017) found a U-shaped curve of TB development, i.e. while young 3-year-olds passed standard TBs, from three-and-a-half years on performances dropped and only until the age of ten years, children began to pass the TBs again. The FB performance, on the other hand, increased with age. Even more astonishing, the study found negative correlations between TB and FB conditions between age three and ten, i.e. those who passed one condition failed the other. An explanation of this paradoxical pattern of findings is that standard TB scenarios create an artificial situation in which everyone (experimenter, protagonist and child) has the same state of knowledge about locations or identities, so that it seems trivial to ask explicitly for a prediction about the protagonist's behavior. In line with this idea of pragmatic confusion rather than a competence limitation, Oktay-Gür and Rakoczy (2017) found that the U-shaped curve of TB performance disappeared once the scenario and the test situation were made less trivial. Applying this logic to the current pattern in the TB conditions suggests that the Duplo task might similarly decrease pragmatic confusion factors compared to standard tasks, leading to a better performance in the Duplo-TB than the standard TB. This

pragmatic advantage does not extend to the FB conditions in the same way because children's competence in understanding false beliefs is still limited (at least in half of our 3-year-old sample). While several studies have found enhanced performance by manipulating different aspects of the standard FB task (e.g., Mitchell & Laco  e, 1991; Psouni et al., 2018; Rhodes & Brandone, 2014; Sullivan & Winner, 1993), meta-analytic findings converge to show that group-level performance remains at chance at 3.5 years of age (Wellman et al., 2001).

The current study failed to reproduce earlier facilitating effects of a simplified FB task, when using matching control conditions. The Duplo task rather seems to tap into the same cognitive system as standard explicit tasks, and make the same demands. Further, correlation and convergence across task types indicate the measurement of a unified ToM competence in 3-year-olds. Yet, the different studies to date using the Duplo task have found very different patterns of results, which leaves open many questions on robustness, replicability and validity of the findings. What we need, in light of the growing amount of recent replication failures in our field, are systematic, large-scale, pre-registered and collaborative cross lab replication studies of measures of early ToM.

Acknowledgment

This study was supported by the German Research Foundation (LI 1989/3-1, RA 2155/4-1, Project: FOR 2253). We are grateful for the participation of all the parents, children and nurseries. Further, we want to thank Fanny Klein for help with data collection, as well as Marlen Kaufmann, Konstanze Schirmer and Jessica Schr  ter for lab coordination. We also want to thank Paula Rubio-Fern  ndez for providing advice on her task.

Study 4: What predicts implicit ToM development?

In this chapter, I will provide a brief overview on a longitudinal investigation of developmental determinants of implicit ToM. We want to thank Marianna Jartó and Johanna Rüter for sharing their data with us and all students who helped in data collection and coding for this study.

Introduction

As reported in the previous chapters of this thesis, we found that measures of infant false belief understanding lack robust replicability and show no convergent validity on the level of concurrent correlations. Thus, these early measures may not represent a unitary ToM capacity. However, one question that remains open is whether implicit mindreading skills exhibit unity on the level of longitudinal correlations. Social interaction accounts on the development of ToM predict that children continuously gain insight into minds by interacting cooperatively with others in joint triadic interactions (e.g., Carpendale & Lewis, 2004; Liszkowski, 2018). On the cognitive side, at the end of the first year of life, infants come to understand others as intentional agents and develop a variety of joint attentional skills, such as the motivation for cooperation and the understanding of others' goals and perspectives, enabling them to interact meaningfully with others in the first place (see e.g., Carpenter et al., 1998; Tomasello & Rakoczy, 2003). Accordingly, many studies to date found later explicit ToM to be predicted by the quality and quantity of early social interactions, such as maternal mind-mindedness, sibling interaction or joint attention (e.g., Licata et al., 2016; McAlister & Peterson, 2007; Meins et al., 2002; Perner et al., 1994; Sodian & Kristen-Antonow, 2015), and by early social understanding (Aschersleben et al., 2008; Wellman et al., 2004). Comparable research on developmental predictors of implicit ToM is still lacking. According to social interaction accounts, implicit ToM, which itself may be precursor of the later explicit form (Thoermer et al., 2012), should be predicted by interactional experience and joint attentional skills as well.

Methods

We piggybacked on an existing longitudinal study that was conducted by researchers in our lab (KOKU Research Center for Cognitive and Cultural Development, University of Hamburg). Amongst other things, this longitudinal study provided us with predictor variables of infants' joint

attentional skills in a point following paradigm (measuring whether children follow the experimenter's point with their gaze at 8 – 12 months; Mundy et al., 2007) and a prosocial helping paradigm (Warneken & Tomasello, 2007). In addition, this study provided us with measures of socio-interactional experience (number of deictic gestures: shows (showing and/or giving an object) and points of infants and parents) from one-hour home visits (HV; recordings of natural everyday interactions at 8, 10 and 18 months), five-minute free play sessions in the lab (FP; child and parent played on the floor with seven toys at 8 – 12 months; Bakeman & Adamson, 1984) and 5-minute decorated room sessions (DR; parent and child walk through a room with 20 interesting objects on the walls at 8 – 14 months; Liskowski & Tomasello, 2011). We re-invited participants at the beginning of their third year of life (median age = 27 months; 10 days, age range = 26;0 – 28;13, 24 included in final sample, 12 boys) and administered as outcome measure a false belief anticipatory correcting paradigm (Knudsen & Liskowski, 2012b) that has recently been replicated successfully (Powell et al., 2018). This interactive task measures whether children spontaneously update an experimenter about the new location of an object that he is about to retrieve (by pointing at the new location) when he is ignorant that the object was swapped (FB condition) compared to when he is knowledgeable (TB condition). Children received two trials of each condition in counterbalanced order (FFTT or TTFF).

Results

Implicit ToM task

In the implicit ToM task, participants tended to point more often to the new object location in the FB condition compared to the TB condition on average across the trials (FB: $M = 1.35$, $SD = 1.87$; TB: $M = 0.83$, $SD = 0.93$; $t(23) = 1.60$, $p = .061$, one-tailed), which is in line with the original study. There were no effects of sex, age or condition order. 67% in the FB condition and 58% in the TB condition informed the experimenter at least once about the new location (contrary to the original study, conditions were not different in this respect; McNemar, $p = .364$, one-tailed). Our findings thus provide a partial replication of the finding by Knudsen & Liskowski (2012).

Correlations with predictor variables

For correlational analyses with predictor variables, we calculated a difference score by subtracting the number of points in the TB condition from the number of points in the FB condition of the ToM task, controlling for a general level of communicativeness (pointing in the TB condition was not necessary). All presented p -values are based on Monte Carlo permutations (10,000 tests). Point following was not predictive of performance in the ToM task. The mean proportion of correct trials in the hidden condition of the helping paradigm at 12 months was correlated with the difference score of the ToM task ($r(22) = .382, p = .077$). Since the helping variable was a relative proportion, we additionally calculated the correlation with a relative variable of ToM performance (pointing in FB condition divided by all points), which yielded even better results ($r(18) = .599, p = .010$). Helping in the hidden condition at 14 months and in the out-of-reach condition were not significantly correlated with performance in the ToM task. The difference score correlated with the number of shows by infants in the FP sessions (9m: $r(24) = .512, p = .017$; 10m: $r(21) = .670, p = .033$; 11m: $r(21) = .401, p = .082$; 12m: $r(24) = .568, p = .012$) and the HV sessions (10m: $r(22) = .553, p = .031$; 18m: $r(22) = .619, p = .017$) at each time point except the earliest at 8 months (there were no shows in the DR). Regarding the number of points by infants (there was barely any pointing during FP), the difference score was correlated only with the HV at 10 months ($r(22) = .540, p = .020$) and the DR at 11 months (especially index finger pointing, $r(20) = .405, p = .077$). Parents' shows and points were not predictive of ToM performance.

Conclusions

In line with social interaction accounts (e.g., Carpendale & Lewis, 2004), the current longitudinal study provides first evidence that early joint attentional skills and socio-interactional experience predict performance in an implicit ToM task, coherent with findings on the development of explicit ToM. In the course of the “nine-month revolution” infants begin to use deictic gestures to communicate outside objects and develop various skills of joint attention (e.g., Carpenter et al., 1998). This special set of new behaviors enables infants to interact meaningfully with others in triadic fashion. Accordingly, our data shows that pointing and showing of objects by infants for their caregiver (measured in the lab and in natural settings at home), and thus sharing attention, facilitated implicit ToM competencies. This suggests that by accumulating experience

in communicating outside entities with others and coordinating perspectives, infants gain insight into the mental world of others. In addition, beside a motivation to cooperate, the hidden condition of the helping paradigm required infants to understand the intention of the agent to get the object but also to take into account her perspective, since the object was out of sight. The absence of a correlation with the out-of-reach condition of the helping task, which only required to understand the agent's need, may indicate that the correlation with the hidden condition was not simply due to a motivation to help. Thus, the correlation of the ToM task with the hidden condition suggests developmental continuity in the understanding of mental states, which indicates unity of early mindreading capacities on the longitudinal level. In the particular case of our interactive ToM measure, the correlations suggest that infants made use of their developing skills of joint attention to solve the task. In the joint activity of searching for the object, infants understood the shared intention to find the object, tracked which information they shared or did not share with the experimenter, and were motivated to update the experimenter when he was missing relevant information. Thus, interactive implicit ToM tasks may not measure a coherent capacity of false belief understanding (Dörrenberg et al., 2018; Poulin-Dubois & Yott, 2018), but a pronounced form of joint attentional skills. See the section on developmental determinants in the general discussion for further elaboration.

General discussion

In the current thesis, I presented four studies dealing with the reliability, convergent validity and origins of implicit ToM abilities. In the following sections, I will summarize our findings and integrate them into a coherent picture. I will also point out issues that we could not answer within the scope of this thesis and provide some future directions.

Summary and synthesis of findings

Reliability of implicit ToM measures

All four studies reported in this thesis provide converging evidence for a lack of robust replicability of implicit measures of FB understanding in young children. None of the different measures that we used (anticipatory looking, looking times and pupil dilation in a violation-of-expectation paradigm, as well as several interaction-based measures such as interactive helping, anticipatory correcting or acting out a narrative) was successfully replicated and none did reveal clear evidence for an early ability to ascribe FBs, as the original studies would suggest.

Anticipatory looking measures performed poorest in revealing FB understanding. Results in study 1 were in the opposite direction than expected by the original findings in the crucial FB2 condition, i.e. in our study, participants anticipated significantly incorrect. The looking pattern of infants clearly suggests that they were not tracking the agent's belief but rather reacted to simpler cues, such as the last location of the object. Our findings are in line with a vast amount of published failed replication studies on this measure from independent labs that mainly appeared during the last year. Several replication studies found negative results using the task by Southgate et al. (2007), the task that we also used (Grosse Wiesmann, Friederici, Disla, Steinbeis, & Singer, 2018; Kulke, Reiß, et al., 2018; Kulke, von Duhn, Schneider, & Rakoczy, 2018; Schuwerk et al., 2018; Zmyj et al., 2015), but also using other anticipatory looking tasks (Burnside, Ruel, Azar, & Poulin-Dubois, 2018; Kulke, von Duhn, et al., 2018) in infants and adults. Crucially, most of these replication studies used larger sample sizes when compared to the original studies (e.g., the original study by Southgate et al. had only 10 infants per test condition), making them less vulnerable to spurious findings. However, the non-replications yielded quite different patterns of results regarding the two test conditions (FB1 and FB2) of the Southgate paradigm. In most studies, participants passed the easier FB1 condition, for which there may be simple

alternative explanations (such as object-tracking). Regarding the more stringent FB2 condition, some studies found the opposite pattern from the original study, and others found chance performance. While the former pattern would favor a non-mentalistic interpretation of the looking behavior, the latter pattern could indicate some sort of tracking knowledge-ignorance of the agent (that she does not know where the object is, instead of having a FB about the location) or even overwhelming demands of the test condition. Thus, in each case, it would be overrated to credit participants with belief-tracking. In light of the amount of FB anticipatory looking studies published to date, a meta-analysis including potential unpublished data might clarify which interpretation most likely is correct.

Our violation-of-expectation measures in study 1 (looking time and pupil dilation) showed some weak indications of infants' sensitivity for others' belief. Only in specific trial orders in within-subject analyses (when the incongruent trial was presented first, but not when the congruent trial was presented first), but not in first trial between-subject analyses (which is contrary to the original studies), we found the expected effects (longer looking/larger pupil in incongruent compared to congruent trials). However, since our task can only be considered a conceptual replication of the paradigm, i.e. we created new stimulus material and used an eye-tracker to measure looking time instead of live coding, we have to be careful with the interpretation. There are certainly other studies that successfully used the violation-of-expectation paradigm to prove early FB understanding (e.g., He et al., 2011; Scott, Baillargeon, Song, & Leslie, 2010). Unfortunately though, most of them come from the same research group (more than 70%, see Poulin-Dubois et al., 2018) and show high variability in the application of inclusion criteria and measurement time without theoretical justification (Rubio-Fernández, 2018a). Recent studies from independent labs that conducted more direct replications of Onishi and Baillargeon's task utterly failed to reproduce the original findings (e.g., Poulin-Dubois & Yott, 2018; Powell et al., 2018). Importantly, even if this paradigm turned out to be reliable, it is particularly prone to simpler explanations than true FB understanding regarding the underlying mindreading abilities, as well as to low-level novelty or other attention capturing effects (see Heyes, 2014a), which emphasizes the necessity for further control conditions in future studies, and tests of convergent validity.

Regarding interaction-based measures of implicit ToM, our studies provide a mixed pattern of findings. Perhaps the most stringent test of FB understanding within the recent set of interactive tasks is the "sefo task" by Southgate et al. (2010), which we targeted in studies 1 and

2. We found a clear preference for choosing the referred box, which in the FB condition is inconsistent with the agent's belief, for children across all age groups and task variants (17-month-olds as in the original study, 2- and 3-year-olds), representing a large sample of about 200 participants. Contrary to the original study, children did not differentiate between true and false belief conditions. In combination with another failed direct replication (Grosse Wiesmann et al., 2017) and a partial conceptual replication (Király et al., 2018), our findings cast doubt on those of the original study. The lack of correlation between the sefo task and a standard FB task at age three suggests different task demands. Surprisingly, the 3-year-olds in our study performed even worse in the sefo compared to the standard task. One of our speculation is that children cannot resist the pointing gesture during the experimenter's request due to a well-known bias to search in pointed-to locations (e.g., Couillard & Woodward, 1999; Palmquist et al., 2012, 2018; Palmquist & Jaswal, 2012; Povinelli & De Blois, 1992). Therefore, the sefo task does not measure -but may even camouflage- the FB competence. How the positive findings in the original study by Southgate et al. were achieved remains unclear. In light of the recent replication failures, it appears likely that the original findings were false-positives.

Our findings from study 3, concerning the narrative act-out FB task, the Duplo task by Rubio-Fernández and Geurts (2013), reject the notion of FB understanding on the group level in 3-year-olds, and suggest that simplified FB tasks show no advantage over standard FB tasks. We failed to replicate an above chance performance in the Duplo task (in line with another failed replication, Kammermeier & Paulus, 2018) and found no performance enhancing effect compared to a closely matched standard task. This corresponds to a recent study by Papafragou et al. (2017) that could not reproduce an advantage of a word learning context on FB understanding, which previous studies would have suggested (Carpenter et al., 2002; Happe & Loth, 2002), when control conditions were closely matched to the test condition. Thus, these studies advise us to carefully inspect findings of earlier FB understanding for the suitability of control conditions, and of course for replicability, before making strong claims about the nature of the underlying ToM capacity.

Another interactive measure, the anticipatory correcting paradigm (Knudsen & Liszkowski, 2012b) that we used in study 4, could be partially replicated. Infants tended to provide more informative points about the new object location in the FB condition compared to the TB condition. Our findings with 2-year-olds are thus in line with the original study that found positive results at 12 and 24 months, and another independent replication study of this paradigm

that was successful at 25 months and 3 years (Powell et al., 2018). However, the anticipatory correcting paradigm may not measure FB understanding. More likely, the paradigm measures whether infants are sensitive to the knowledge or ignorance of the agent (e.g., he does not know where the object is, or he does not know that the object was transferred), rather than to a misrepresentation (e.g., he thinks the object is in the other box). Based on infants' understanding of the agent's goal to achieve the object, they can inform appropriately as soon as the agent lacks information. The same logic can be applied to another interactive ToM task that is quite similar to the anticipatory correcting paradigm but measures backward-processing of an agent's behavior rather than anticipation of mistakes, the Buttelmann task (D. Buttelmann et al., 2009). In the Buttelmann task, infants helped an experimenter that tried to open an empty box by opening the box containing the object, when he held a FB about the object location. Without really attributing a FB, infants can pass the task by using their understanding of the agent's ignorance about the object location and of his goal to achieve the object. In the corresponding TB condition, infants helped to open the empty box. This suggests that they attributed the agent in the TB condition the goal to open the empty box (since he was aware of the current object location). Interestingly, as for the anticipatory correcting paradigm, several published studies successfully or at least partially successfully replicated the task by Buttelmann et al. (Fizke et al., 2017; Oktay-Gür et al., 2018; Powell et al., 2018; Priewasser, Rafetseder, Gargitter, & Perner, 2018). Priewasser et al. (2018) added another control condition to the Buttelmann task, in which the agent in the FB condition tried to open a third box that was always empty (neutralizing the agent's FB). Yet, although that new box was not matching the agent's FB, infants helped the agent by opening the box containing the object. This finding indicates that infants in fact rather tracked the agent's ignorance than FB.

Taken together, none of the various measures of early FB understanding turned out to be reliably replicable in our studies, as well as in a bunch of other recent replication studies. This challenges the original studies and the existence of early FB representation per se. In fact, our and others' replication studies suggest that early interactive measures of knowledge-ignorance (e.g., D. Buttelmann et al., 2009; Knudsen & Liszkowski, 2012b) are more promising regarding replicability compared to measures of FB (e.g., Southgate et al., 2010, 2007). Thus, here, at the understanding of someone else's knowledge state, may be the true borderline of infants' meta-representational capacities.

Convergent validity

Let us assume the different implicit ToM tasks turned out to be reliable. Reliability, however, would not settle the issue of validity. One important question regarding implicit tasks is whether they all actually measure the same underlying competence, namely implicit ToM according to two-systems views (e.g., Apperly & Butterfill, 2009), or even proper ToM according to nativist views (e.g., Scott & Baillargeon, 2017). If this would be the case, performances between different tasks should show correlation, which would constitute convergent validity, as amply documented for explicit ToM (Gopnik & Astington, 1988; Perner & Roessler, 2012; Rakoczy et al., 2015). Accordingly, in study 1, we correlated performances between three of the main paradigms of early FB understanding: Violation-of-expectation (Onishi & Baillargeon, 2005), anticipatory looking (Southgate et al., 2007) and interactive helping (Southgate et al., 2010). Unfortunately, we found no correlations among the three different paradigms. Thus, different implicit ToM tasks do not measure a coherent and unitary social skill, such as FB understanding. Each task may measure a different capacity. Only for speculation: Violation-of-expectation tasks may measure low-level novelty, anticipatory looking tasks may measure intention-tracking (without taking into account the belief of the agent), and interactive tasks may measure knowledge-ignorance (not FB). It could also be that the tasks show no correlations because they demand different cognitive loads. However, all these implicit ToM tasks were specifically designed to reduce processing load and were mastered by the majority of infants in the original studies, thus, differences in task demands should be minimal.

For the narrative act-out FB task (Duplo task; Rubio-Fernández & Geurts, 2013) used in study 3, we found a different picture. Performances of the Duplo task, a standard FB task and an intentionality version of the Duplo task showed convergence and correlation. Thus, although we found no advantage of the Duplo task on FB performance, our findings indicate that this task measures the same capacity as the standard task, namely explicit ToM. The correlation to the intentionality version further indicates that some 3-year-olds (they performed at chance on the group level) already have a flexible, full-blown ToM competence comparable to that of older children. In sum, these findings support the classical view on ToM development that children fail in FB tasks until age three and show a gradual increase in performance on the group level between three and four years of age (e.g., Wellman et al., 2001).

Only recently, several other studies targeted the question of convergent validity of implicit ToM tasks and found the same negative results across the board. One study found no

correlation between an anticipatory looking and an interactive task (Grosse Wiesmann et al., 2017), two studies found no correlation between violation-of-expectation and interactive tasks (Poulin-Dubois & Yott, 2018; Powell et al., 2018), and two more studies found no correlation between different anticipatory looking tasks (Kulke, Reiß, et al., 2018; Kulke, von Duhn, et al., 2018). The lack of convergent validity of implicit ToM tasks rejects the notion that these tasks represent cumulative evidence for an early and automatic ToM capacity as proposed by two-systems accounts, as well as for an inherent proper ToM capacity as proposed by nativists. Much more likely are views of skeptics that assume alternative explanations for each local finding, such as behavior reading or the ascription of simpler mental states (e.g., Heyes, 2014a). However, in combination with the lack of reliability of implicit ToM tasks, any theoretical claim seems premature.

Developmental determinants

In study 4, we found that joint attentional behavior in the first year of life (as early as 9 months), such as using deictic gestures (pointing, showing), predicted performance in an interactive ToM task at two years. Importantly, this correlation between joint attention and implicit ToM was not caused by the general level of communicativeness (that those who showed more deictic gestures early in life simply showed more deictic gestures in our ToM task) but by meaningful informative communication about mental states. That is, we used pointing in the TB condition (where it was unnecessary but still frequently occurred) as a baseline for communicativeness and calculated a difference score of FB performance as outcome variable. Accordingly, those participants who engaged more frequently in joint triadic interaction in their first year, more likely took into account the ignorance of the experimenter at two years. In addition, we found evidence that implicit ToM builds on an early understanding of simpler mental states such as seeing or intending. That is, implicit ToM correlated with the hidden condition of the helping paradigm. This task required understanding not only the agent's intention to obtain the object, but also the different perspective of the agent since his view of an object was blocked. This provides further evidence for a continuous development of early mindreading skills: From the understanding of simpler mental states such as seeing, to more complex mental states such as knowing. Our findings thus indicate developmental continuity in the social domain and an important, maybe foundational, role of triadic interactions in acquiring mental state concepts. This is in line with social interaction accounts on the development of ToM

(e.g., Carpendale & Lewis, 2004), and converges with findings showing that explicit FB understanding equally builds on joint attention and social understanding (e.g., Sodian & Kristen-Antonow, 2015; Wellman et al., 2008). The latter suggests a unitary developmental pathway of implicit and explicit ToM that might lead continuously from one to the other (Thoermer et al., 2012).

Although we conducted study 4 in a longitudinal design, it remains possible that the correlation between joint attention and implicit ToM could also be found in a cross-sectional investigation at an early time point, and that joint attention is not a developmental foundation of early mindreading skills. As discussed in the previous sections, the anticipatory correcting task used as outcome measure in our longitudinal study may not measure FB understanding but rather knowledge-ignorance. Several studies suggest that infants comprehend what others see and know quite early in development (from around 12 months on; Liskowski, Carpenter, & Tomasello, 2008; Tomasello & Haberl, 2003), and even great apes have this competence (e.g., Hare, Call, & Tomasello, 2001). Yet, an argument against a cross-sectional correlation at an early time point might be that we found joint attention as a predictor of implicit ToM already from nine months on, when infants may not yet have an understanding of knowledge-ignorance. However, a recent shared intentionality account that integrated early findings of implicit ToM suggests that infants and apes use their early competence of tracking epistemic states to solve implicit ToM tasks (Tomasello, 2018). Apes, however, do not engage in joint attention (e.g., Tomasello, Carpenter, Call, Behne, & Moll, 2005). Then how can we integrate our finding of early joint attention as a predictor of implicit ToM into such an account? One possibility might be that joint attention does not constitute a prerequisite to the development of epistemic state understanding that is necessary to predict behavior in implicit ToM tasks. Joint attention might rather be a facilitator of the processes involved by providing an environment for learning about different perspectives. Another possibility might be that apes follow a different ontogenetic trajectory to establish a similar understanding of epistemic states as humans. However, it should be noted that findings on apes' competence in implicit ToM tasks have not been investigated with regard to their reliability and convergent validity, yet. Thus, a foundational role of joint attention in acquiring those skills measured by implicit ToM tasks remains a possibility.

Obviously, the correlation between joint attention and implicit ToM can be interpreted in several ways. On the one hand, joint attention implies triadic interaction and thus interactive experience. When infants show objects to their parents or point at objects, parents usually

respond and engage in conversation about the object. This provides learning opportunities involving the coordination of different perspectives and epistemic states. On the other hand, joint attention behavior could be seen as an acted out form of the underlying socio-cognitive capacities of mental state understanding. In the latter case, the correlation with implicit ToM would rather imply a continuous development of mindreading skills, i.e. early “simpler” mindreading (represented in joint attention behavior) that predicts later “more complex” mindreading (implicit ToM), though it would not provide conclusive evidence that infants gain insight into other minds via social interactions. It is a “chicken or the egg” causality dilemma: Do infants engage in joint attention because they understand something about mental states, or do they understand something about mental states because they interact with others? Our findings cannot provide an ultimate answer to that issue, but it is probably a combination of both. Early joint attention behaviors at 12 months have been shown to indicate some sort of mental state understanding (e.g., Liszkowski et al., 2008). Though, recent longitudinal and cross-cultural studies suggest that early social interaction may be crucial for the emergence of joint attention behaviors in the first place (see for review Liszkowski, 2018). For instance, it has been shown that infants of parents who point more start pointing earlier and point more themselves (Liszkowski, Brown, Callaghan, Takada, & de Vos, 2012; Liszkowski & Tomasello, 2011). In addition, a recent study found that infants’ pointing frequency at 12 months was predicted by the level of parental responsiveness at 10 months (Ger, Altınok, Liszkowski, & Küntay, 2018) suggesting an important role of shared reference (i.e., by pointing infants “align their and others’ cognitive engagement in the world”, Liszkowski, 2018, p. 26) in the development of joint attention behaviors. Thus, in triadic interactions with caregivers, infants get a first grasp of other’s perspectives and understand that others may or may not share information with them. By using this understanding in joint engagements over time, infants continuously gain more complex mindreading concepts such as the understanding that others’ perspective may diverge from their own (explicit ToM; Sodian & Kristen-Antonow, 2015).

Limitations and outlook

The studies presented here provide an important contribution to infant ToM research. Yet, the jury is still out. Each of our studies may have methodological limitations, there may be post-hoc

alternative explanations for replication failures or lack of correlations, and thus limits in the validity of our findings.

For instance, contrary to most of the original studies, in some of our studies we used within-subject designs and administered multiple tasks and trials. Although this was a necessary step for comparison of task performances and for calculating correlations between tasks, it comes with a potential danger of carry-over effects from one task to the other. In most cases, we could control for these effects by counterbalancing task order (e.g., studies 2 and 3). Although we kept task order constant in study 1 to avoid irrelevant variance, we took precautions to minimize potential carry-over effects. That is, since we conducted two eye-tracking tasks in study 1, we administered the one with lengthy stimulus videos and belief-inconsistent action outcomes (violation-of-expectation task) in the end to avoid fatigue and carry-over effects on subsequent tasks. In addition, we separated the eye-tracking tasks with the interactive task in the middle to include a natural break with playful elements. It should be noted, however, that there are other published replication studies that used multiple implicit tasks but controlled for task order and were still unsuccessful in finding original effects (e.g., Poulin-Dubois & Yott, 2018). Carry-over effects could not explain the weak performance in the anticipatory looking task in study 1 anyway, since this task was always conducted at the first position. In addition, in study 2, we revisited the interactive sefo task (that was always the second task in study 1) using a single task design, and yet found the same negative results, which argues against negative impacts due to the design of our study 1.

Another important issue is the methodological accuracy of replication studies. Subtle variations of the original study's protocol could potentially lead to worse performance or may even ruin original effects (see for discussion Baillargeon et al., 2018; Poulin-Dubois et al., 2018). This may especially pertain to interactive and violation-of-expectation tasks, and less to anticipatory looking tasks. Anticipatory looking tasks can be conducted in automated fashion by using an eye-tracker to measure gaze behavior and by presenting the original stimulus videos. Thus, results of replication studies of anticipation tasks appear quite solid. Interactive tasks and violation-of-expectation tasks, on the other hand, mostly require intensive investigation of experimental materials, procedure and coding scheme, and require sophisticated acting from the experimenter. This leaves plenty of room for accidental deviations from the original methods. Unfortunately, method sections of papers are often kept short, are in parts ambiguous, and do often not provide enough details and comprehensive information to exactly replicate

experiments. In some cases, video materials of test sessions may not even exist. These factors complicate the implementation of a direct replication and leave room for speculations about potential alternative explanations for replication failures. Although we contacted the original authors to get advice on their tasks, each of our implementations certainly differed in minor aspects from the original experiments (e.g., precise age range, warm-up routine, light conditions, movement speed of experimenter, and so on). However, since a variety of these potentially distracting factors were not theoretically motivated in the original studies either, the question whether or not deviations are in fact responsible for replication failures should be targeted in controlled empirical manner or via meta-analytic approaches.

Regarding our approach of testing for convergent validity, one may argue that correlations between tasks would be less likely expected when the original effects were not replicated in the first place. Although this may sound convincing, it is still possible that some infants reveal their mindreading skills in failed replication tasks, even if participants as a group fail. That is, chance performance could be interpreted as random choosing, though it could also be the case that half of the participants pass not accidentally but because they have the competence. This is actually what we found in study 3 with 3-year-olds in the act-out FB task. Accordingly, if, for instance, only half of the infants pass several tasks, and their performance would reveal their implicit ToM competence in these tasks, correlations should still be expected. If task performances were utterly random or due to very different underlying phenomena, of course, there should be no inter-task correlations.

Additionally, to be fair, a single failed replication of a single task cannot stagger a substantial body of evidence such as 30 studies on implicit ToM from various measures (see Scott & Baillargeon, 2017). However, this is not the case. To date, our negative findings on implicit ToM converge with a variety of other published negative findings from independent labs. These failed replication studies itself cover a variety of measures and fail to find evidence for convergent validity across the board. In addition, there is a potential body of unpublished failed replication attempts in the file-drawers (Kulke & Rakoczy, 2018). Importantly, the number of tested participants in most of the published replication studies resembles or exceeds those of the original studies, and in combination, current replication studies exceed some of the original sample sizes by far. The fact that there are so many published positive findings of early FB understanding, thus, could be merely a result of a publication bias. Accordingly, it would be inappropriate to attribute original studies more credibility than replication studies simply

because they were published first. In sum, the evidential status of implicit ToM research is quite unclear: We currently find for almost each original measure one or more failed replication attempts. Therefore, we should take this replication crisis seriously and assume that we currently do not know whether children have FB understanding before they pass explicit ToM tasks at four years of age.

However, as already correctly claimed in this debate: “Absence of evidence is, of course, no evidence of absence” (Poulin-Dubois et al., 2018, p. 307). In this respect, a next step would be to jointly target the issue of replicability in our field. On the one hand, meta-analyses of published and unpublished data on implicit ToM can inform us about the current evidential status of effects and unravel certain governing factors such as participant age, warm-up routines, inclusion criteria, or other methodological variations of studies. On the other hand, multi-lab, collaborative conceptual replication attempts of individual tasks can provide independent and widely acceptable consensus on the existence of effects. Such large-scale replications would offer the opportunity to test a variety of influencing factors and control conditions, unraveling the scopes and limits of (possible) implicit ToM capacities. They would also allow to conduct tests of convergent validity with agreed-on versions and dependent measures of implicit ToM tasks. In fact, such a collaborative replication project is already deep in the planning and piloting stage. In a “ManyBabies” project on implicit ToM (see Frank et al., 2017) researchers from labs all over the world (including some of the authors of the original studies) currently implement tasks together using anticipatory looking, interactive and violation-expectation measures. Thus, the next months and years will hopefully clarify whether there is such a thing as implicit ToM and whether infants can really attribute false beliefs.

Concluding remarks

Our findings advise a pessimistic stance towards early FB understanding and suggest that strong theoretical claims were raised prematurely. In fact, our findings argue for a more classical view on the development of ToM: Children between one and three years of age do not systematically pass FB tasks. Yet, robust empirical evidence for the opposite view and successful cross-validations are still outstanding. In the meantime, it looks as if infants’ mindreading abilities are limited to tasks involving lower representational demands, such as the knowledge-ignorance distinction. Intriguingly, infants’ competence in epistemic state understanding seems to be

rooted in simpler social understanding and socio-interactional experience. Thus, our findings speak for a continuous development of mindreading abilities towards later FB understanding. Some fundamental conceptual changes seem to occur during the fourth year of life when some kids already reveal sophisticated ToM skills, though not yet on the group level. As classical findings on explicit ToM suggest, these changes in the understanding of different and diverging perspectives may build on earlier social skills, interactive and linguistic experience, as well as general cognitive maturation in executive control.

To date, there is no indication that implicit ToM is a real phenomenon (at least not in young children; but see for doubts about implicit ToM in adults, e.g., Phillips et al., 2015; Santiesteban, Shah, White, Bird, & Heyes, 2015). The amount and diversity of failed replication studies and failed studies of convergent validity suggest that original findings may constitute false-positives. The driving force of this replication crisis is likely a publication bias. Due to journal publishing policies, mainly studies presenting statistically significant data and novel contributions to the scientific field make it into press, thus barely any replications, especially not unsuccessful ones. As a consequence, for each published positive finding, there may be dozens of unpublished negative findings. First evidence for a publication bias in our field of infant ToM research was provided by a recent survey (Kulke & Rakoczy, 2018) and a special issue in the journal *Cognitive Development* on replication attempts (Sabbagh & Paulus, 2018; where most published failed replications including ours arose from). The current replication crisis calls for an increasing importance for future studies to pre-register hypotheses and study designs, as well as for an increasing awareness of open science, to maintain credibility and relevance of our field.

References

- Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. <https://doi.org/10.1037/a0016923>
- Apperly, I. A., & Robinson, E. J. (1998). Children's mental representation of referential relations. *Cognition*, *67*, 287–309. [https://doi.org/10.1016/S0010-0277\(98\)00030-4](https://doi.org/10.1016/S0010-0277(98)00030-4)
- Aschersleben, G., Hofer, T., & Jovanovic, B. (2008). The link between infant attention to goal-directed action and later theory of mind abilities. *Developmental Science*, *11*(6), 862–868. <https://doi.org/10.1111/j.1467-7687.2008.00736.x>
- Astington, J. W., & Gopnik, A. (1988). *Knowing you've changed your mind: Children's understanding of representational change*. New York: Cambridge University Press.
- Baillargeon, R. (1987). Object permanence in 3½- and 4½-month-old infants. *Developmental Psychology*, *23*(5), 655–664. <https://doi.org/10.1037/0012-1649.23.5.655>
- Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited Commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, *46*, 112–124. <https://doi.org/10.1016/j.cogdev.2018.06.001>
- Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology*, *67*(1), 159–186. <https://doi.org/10.1146/annurev-psych-010213-115033>
- Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118. <https://doi.org/10.1016/j.tics.2009.12.006>
- Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208. [https://doi.org/10.1016/0010-0277\(85\)90008-3](https://doi.org/10.1016/0010-0277(85)90008-3)
- Bakeman, R., & Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, *55*(4), 1278–89.
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. <https://doi.org/10.1177/1745691612459060>
- Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the

- communicative intentions behind gestures in a hiding game. *Developmental Science*, 8(6), 492–499. <https://doi.org/10.1111/j.1467-7687.2005.00440.x>
- Behne, T., Liszowski, U., Carpenter, M., & Tomasello, M. (2012). Twelve-month-olds' comprehension and production of pointing. *British Journal of Developmental Psychology*, 30(3), 359–375. <https://doi.org/10.1111/j.2044-835X.2011.02043.x>
- Bennett, J. (1978). Some remarks about concepts. *The Behavioral and Brain Sciences*, 4, 557–560. <https://doi.org/10.1017/S0140525X00076573>
- Bialecka-Pikul, M., Kosno, M., Białek, A., & Szpak, M. (2019). Let's do it together! The role of interaction in false belief understanding. *Journal of Experimental Child Psychology*, 177, 141–151. <https://doi.org/10.1016/j.jecp.2018.07.018>
- Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2018). Implicit false belief across the lifespan: Non-replication of an anticipatory looking task. *Cognitive Development*, 46, 4–11. <https://doi.org/10.1016/j.cogdev.2017.08.006>
- Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PLoS ONE*, 12(4), 1–13. <https://doi.org/10.1371/journal.pone.0173793>
- Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112(2), 337–342. <https://doi.org/10.1016/j.cognition.2009.05.006>
- Buttelmann, D., Over, H., Carpenter, M., & Tomasello, M. (2014). Eighteen-month-olds understand false beliefs in an unexpected-contents task. *Journal of Experimental Child Psychology*, 119, 120–126. <https://doi.org/10.1016/j.jecp.2013.10.002>
- Buttelmann, F., Suhrke, J., & Buttelmann, D. (2015). What you get is what you believe: Eighteen-month-olds demonstrate belief understanding in an unexpected-identity task. *Journal of Experimental Child Psychology*, 131, 94–103. <https://doi.org/10.1016/j.jecp.2014.11.009>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. <https://doi.org/10.1038/nrn3475>

- Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5), 187–192. <https://doi.org/10.1016/j.tics.2008.02.010>
- Carlson, S. M., & Moses, L. J. (2001). Individual Differences in Inhibitory Control and Children's Theory of Mind. *Child Development*, 72(4), 1032–1053. <https://doi.org/10.1111/1467-8624.00333>
- Carpendale, J. I. M., & Lewis, C. (2004). Constructing an understanding of mind: The development of children's social understanding within social interaction. *Behavioral and Brain Sciences*, 27, 79-96; discussion 96-151. <https://doi.org/10.1017/S0140525X04000032>
- Carpenter, M., Call, J., & Tomasello, M. (2002). A new false belief test for 36-month-olds. *British Journal of Developmental Psychology*, 20(3), 393–420. <https://doi.org/10.1348/026151002320620316>
- Carpenter, M., Nagell, K., Tomasello, M., & Butterworth, G. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs for the Society for Research in Child Development*, 63(4), 179. <https://doi.org/10.2307/1166214>
- Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, 28(2), 141–172. <https://doi.org/10.1111/mila.12014>
- Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, 9(4), 377–395. [https://doi.org/10.1016/0885-2014\(94\)90012-4](https://doi.org/10.1016/0885-2014(94)90012-4)
- Couillard, N. L., & Woodward, A. L. (1999). Children's comprehension of deceptive points. *British Journal of Developmental Psychology*, 17(4), 515–521. <https://doi.org/10.1348/026151099165447>
- Crivello, C., & Poulin-Dubois, D. (2018). Infants' false belief understanding: A non-replication of the helping task. *Cognitive Development*, 46, 51–57. <https://doi.org/10.1016/j.cogdev.2017.10.003>
- Davis, H. L., & Pratt, C. (1995). The development of children's theory of mind: The working memory explanation. *Australian Journal of Psychology*, 47(1), 25–31. <https://doi.org/10.1080/00049539508258765>

- Dennett, D. C. (1978). Beliefs about beliefs. *The Behavioral and Brain Sciences*, 4, 568–570. <https://doi.org/https://doi.org/10.1017/S0140525X00076664>
- Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, 46, 12–30. <https://doi.org/10.1016/j.cogdev.2018.01.001>
- Dörrenberg, S., Wenzel, L., Proft, M., Rakoczy, H., & Liszkowski, U. (2019). Reliability and generalizability of an acted-out false belief task in 3-year-olds. *Infant Behavior and Development*, 54, 13–21. <https://doi.org/10.1016/j.infbeh.2018.11.005>
- Elsner, B., Pauen, S., & Jeschonek, S. (2006). Physiological and behavioral parameters of infants' categorization: Changes in heart rate and duration of examining across trials. *Developmental Science*, 9(6), 551–556. <https://doi.org/10.1111/j.1467-7687.2006.00532.x>
- Fabricius, W. V., Boyer, T. W., Weimer, A. A., & Carroll, K. (2010). True or false: Do 5-year-olds understand belief? *Developmental Psychology*, 46(6), 1402–1416. <https://doi.org/10.1037/a0017648>
- Falck-Ytter, T. (2008). Face inversion effects in autism: A combined looking time and pupillometric study. *Autism Research*, 1(5), 297–306. <https://doi.org/10.1002/aur.45>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Fizke, E., Butterfill, S., van de Loo, L., Reindl, E., & Rakoczy, H. (2017). Are there signature limits in early Theory of Mind? *Journal of Experimental Child Psychology*, 162, 209–224. <https://doi.org/10.1016/j.jecp.2017.05.005>
- Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, 22(4), 421–435. <https://doi.org/10.1111/inf.12182>
- Ger, E., Altınok, N., Liszkowski, U., & Küntay, A. C. (2018). Development of infant pointing from 10 to 12 months: The role of relevant caregiver responsiveness. *Infancy*, 23(5), 708–729. <https://doi.org/10.1111/inf.12239>

- Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*(6873), 755. <https://doi.org/10.1038/415755a>
- Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193. [https://doi.org/10.1016/0010-0277\(95\)00661-H](https://doi.org/10.1016/0010-0277(95)00661-H)
- Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, *59*(1), 26–37. <https://doi.org/10.2307/1130386>
- Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind* (pp. 257–293). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511752902.011>
- Gredebäck, G., Eriksson, M., Schmitow, C., Laeng, B., & Stenberg, G. (2012). Individual differences in face processing: Infants' scanning patterns and pupil dilations are influenced by the distribution of parental leave. *Infancy*, *17*(1), 79–101. <https://doi.org/10.1111/j.1532-7078.2011.00091.x>
- Gredebäck, G., & Melinder, A. (2011). Teleological reasoning in 4-month-old infants: Pupil dilations and contextual constraints. *PLoS ONE*, *6*(10), e26487. <https://doi.org/10.1371/journal.pone.0026487>
- Grosse Wiesmann, C., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2018). Longitudinal evidence for 4-year-olds' but not 2- and 3-year-olds' false belief-related action anticipation. *Cognitive Development*, *46*, 58–68. <https://doi.org/10.1016/j.cogdev.2017.08.007>
- Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children. *Developmental Science*, *20*(5), e12445. <https://doi.org/10.1111/desc.12445>
- Hamilton, A. F. D. C., Brindley, R., & Frith, U. (2009). Visual perspective taking impairment in children with autistic spectrum disorder. *Cognition*, *113*(1), 37–44. <https://doi.org/10.1016/j.cognition.2009.07.007>
- Happe, F., & Loth, E. (2002). "Theory of Mind" and tracking speakers' intentions. *Mind and Language*, *17*(1&2), 24–36. <https://doi.org/10.1111/1468-0017.00187>

- Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*(1), 139–151. <https://doi.org/10.1006/anbe.2000.1518>
- He, Z., Bolz, M., & Baillargeon, R. (2011). False-belief understanding in 2.5-year-olds: Evidence from violation-of-expectation change-of-location and unexpected-contents tasks. *Developmental Science*, *14*(2), 292–305. <https://doi.org/10.1111/j.1467-7687.2010.00980.x>
- Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences*, *18*(4), 167–170. <https://doi.org/10.1016/j.tics.2014.01.005>
- Hepach, R., Vaish, A., & Tomasello, M. (2012). Young children are intrinsically motivated to see others helped. *Psychological Science*, *23*(9), 967–972. <https://doi.org/10.1177/0956797612440571>
- Hepach, R., & Westermann, G. (2016). Pupillometry in infancy research. *Journal of Cognition and Development*, *17*(3), 359–377. <https://doi.org/10.1080/15248372.2015.1135801>
- Heyes, C. (2014a). False belief in infancy: A fresh look. *Developmental Science*, *17*(5), 647–659. <https://doi.org/10.1111/desc.12148>
- Heyes, C. (2014b). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*(2), 131–143. <https://doi.org/10.1177/1745691613518076>
- Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, *57*(3), 567–582. <https://doi.org/10.2307/1130337>
- Holmes, H. A., Black, C., & Miller, S. A. (1996). A cross-task comparison of false belief understanding in a head start population. *Journal of Experimental Child Psychology*, *63*(2), 263–285. <https://doi.org/10.1006/jecp.1996.0050>
- Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science*, *12*(4), 670–679. <https://doi.org/10.1111/j.1467-7687.2008.00805.x>
- Kammermeier, M., & Paulus, M. (2018). Do action-based tasks evidence false-belief understanding in young children? *Cognitive Development*, *46*, 31–39.

<https://doi.org/10.1016/j.cogdev.2017.11.004>

- Király, I., Oláh, K., Csibra, G., & Kovács, Á. M. (2018). Retrospective attribution of false beliefs in 3-year-old children. *Proceedings of the National Academy of Sciences*, *115*(45), 11477–11482. <https://doi.org/doi.org/10.1073/pnas.1803505115>
- Király, I., Oláh, K., Kovács, Á., & Csibra, G. (2016). Do 18- and 36-month-old infants update attributed beliefs by re-evaluating past events? *Poster Session Presented at the Budapest CEU Conference on Cognitive Development, Budapest, Hungary.*
- Knudsen, B., & Liszkowski, U. (2012a). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*, *17*(6), 672–691. <https://doi.org/10.1111/j.1532-7078.2011.00105.x>
- Knudsen, B., & Liszkowski, U. (2012b). Eighteen- and 24-month-old infants correct others in anticipation of action mistakes. *Developmental Science*, *15*(1), 113–122. <https://doi.org/10.1111/j.1467-7687.2011.01098.x>
- Köster, M., Ohmer, X., Nguyen, T. D., & Kärtner, J. (2016). Infants understand others' needs. *Psychological Science*, *27*(4), 542–548. <https://doi.org/10.1177/0956797615627426>
- Kovács, A. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–1834. <https://doi.org/10.1126/science.1190792>
- Kovács, A. M., Téglás, E., & Endress, A. D. (2016). Automatic belief tracking effects cannot be explained by attention check timing: Reply to Phillips et al. *Unpublished Manuscript.*
- Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114. <https://doi.org/10.1126/science.aaf8110>
- Kulke, L., & Rakoczy, H. (2018). Implicit Theory of Mind – An overview of current replications and non-replications. *Data in Brief*, *16*, 101–104. <https://doi.org/10.1016/j.dib.2017.11.016>
- Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development*, *46*, 97–111. <https://doi.org/10.1016/j.cogdev.2017.09.001>

- Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit Theory of Mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, 29(6), 888–900. <https://doi.org/10.1177/0956797617747090>
- Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, 9(10), 459–462. <https://doi.org/10.1016/j.tics.2005.08.006>
- Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in “Theory of Mind.” *Trends in Cognitive Sciences*, 8(12), 528–533. <https://doi.org/10.1016/j.tics.2004.10.001>
- Licata, M., Kristen, S., & Sodian, B. (2016). Mother-Child Interaction as a Cradle of Theory of Mind: The Role of Maternal Emotional Availability. *Social Development*, 25(1), 139–156. <https://doi.org/10.1111/sode.12131>
- Liszkowski, U. (2018). Emergence of shared reference and shared minds in infancy. *Current Opinion in Psychology*, 23, 26–29. <https://doi.org/10.1016/j.copsyc.2017.11.003>
- Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & de Vos, C. (2012). A prelinguistic gestural universal of human communication. *Cognitive Science*, 36(4), 698–713. <https://doi.org/10.1111/j.1551-6709.2011.01228.x>
- Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, 108(3), 732–739. <https://doi.org/10.1016/j.cognition.2008.06.013>
- Liszkowski, U., & Tomasello, M. (2011). Individual differences in social, cognitive, and morphological aspects of infant pointing. *Cognitive Development*, 26(1), 16–29. <https://doi.org/10.1016/j.cogdev.2010.10.001>
- Lohmann, H., Carpenter, M., & Call, J. (2005). Guessing versus choosing - and seeing versus believing - in false belief tasks. *British Journal of Developmental Psychology*, 23(3), 451–469. <https://doi.org/10.1348/026151005X26877>
- Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: a training study. *Child Development*, 74(4), 1130–1144.
- Low, J. (2010). Preschoolers’ implicit and explicit false belief understanding: Relations with complex syntactic mastery. *Child Development*, 81(2), 597–615.

- Low, J., Apperly, I. A., Butterfill, S. A., & Rakoczy, H. (2016). Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account. *Child Development Perspectives, 10*(3), 184–189. <https://doi.org/10.1111/cdep.12183>
- Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science, 24*(3), 305–311. <https://doi.org/10.1177/0956797612451469>
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542. <https://doi.org/10.1177/1745691612460688>
- McAlister, A., & Peterson, C. (2007). A longitudinal study of child siblings and theory of mind development. *Cognitive Development, 22*(2), 258–270. <https://doi.org/10.1016/j.cogdev.2006.10.009>
- Meins, E., Fernyhough, C., Wainwright, R., Das Gupta, M., Fradley, E., & Tuckey, M. (2002). Maternal mind-mindedness and attachment security as predictors of theory of mind understanding. *Child Development, 73*(6), 1715–1726. <https://doi.org/10.1111/1467-8624.00501>
- Meltzoff, A. N., & Gopnik, A. (1993). The role of imitation in understanding persons and developing a theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from autism* (pp. 335–366). New York: Oxford University Press.
- Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L., & Siegal, M. (2012). Belief attribution in deaf and hearing infants. *Developmental Science, 15*(5), 633–640. <https://doi.org/10.1111/j.1467-7687.2012.01155.x>
- Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*(2), 622–646. <https://doi.org/10.1111/j.1467-8624.2007.01018.x>
- Mitchell, P., & Lacohee, H. (1991). Children's early understanding of false belief. *Cognition, 39*, 107–127.
- Moll, H., Kane, S., & McGowan, L. (2016). Three-year-olds express suspense when an agent

- approaches a scene with a false belief. *Developmental Science*, *19*(2), 208–220.
<https://doi.org/10.1111/desc.12310>
- Moll, H., Khalulyan, A., & Moffett, L. (2017). 2.5-year-olds express suspense when others approach reality with false expectations. *Child Development*, *88*(1), 114–122.
<https://doi.org/10.1111/cdev.12581>
- Moll, H., & Tomasello, M. (2004). 12- and 18-month-old infants follow gaze to spaces behind barriers. *Developmental Science*, *7*(1), F1–F9. <https://doi.org/10.1111/j.1467-7687.2004.00315.x>
- Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., & Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child Development*, *78*(3), 938–954. <https://doi.org/10.1111/j.1467-8624.2007.01042.x>
- Oktay-Gür, N., & Rakoczy, H. (2017). Children’s difficulty with true belief tasks: Competence deficit or performance problem? *Cognition*, *166*, 28–41.
<https://doi.org/10.1016/j.cognition.2017.05.002>
- Oktay-Gür, N., Schulz, A., & Rakoczy, H. (2018). Children exhibit different performance patterns in explicit and implicit Theory of Mind tasks. *Cognition*, *173*, 60–74.
<https://doi.org/10.1016/j.cognition.2018.01.001>
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. <https://doi.org/10.1126/science.1107621>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Palmquist, C. M., Burns, H. E., & Jaswal, V. K. (2012). Pointing disrupts preschoolers’ ability to discriminate between knowledgeable and ignorant informants. *Cognitive Development*, *27*(1), 54–63. <https://doi.org/10.1016/j.cogdev.2011.07.002>
- Palmquist, C. M., & Jaswal, V. K. (2012). Preschoolers expect pointers (even ignorant ones) to be knowledgeable. *Psychological Science*, *23*(3), 230–231.
<https://doi.org/10.1177/0956797611427043>
- Palmquist, C. M., Kondrad, R. L., & Norris, M. N. (2018). Follow my point? Preschoolers’

- expectations about veridicality disrupt their understanding of deceptive points. *Cognitive Development*, 48(January), 190–202. <https://doi.org/10.1016/j.cogdev.2018.08.009>
- Papafragou, A., Fairchild, S., Cohen, M. L., & Friedberg, C. (2017). Learning words from speakers with false beliefs. *Journal of Child Language*, 44(4), 905–923. <https://doi.org/10.1017/S0305000916000301>
- Paulus, M., Hunnius, S., van Wijngaarden, C., Vrins, S., van Rooij, I., & Bekkering, H. (2011). The role of frequency information and teleological reasoning in infants' and adults' action prediction. *Developmental Psychology*, 47(4), 976–983. <https://doi.org/10.1037/a0023785>
- Perner, J. (2014). Commentary on Ted Ruffman's "Belief or not belief:" *Developmental Review*, 34(3), 294–299. <https://doi.org/10.1016/j.dr.2014.05.002>
- Perner, J., Huemer, M., & Leahy, B. (2015). Mental files in development: A cognitive theory of how children represent belief and its intensionality. *Cognition*, 145, 77–88. <https://doi.org/10.1016/j.cognition.2015.08.006>
- Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences*, 16(10), 519–525. <https://doi.org/10.1016/j.tics.2012.08.004>
- Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, 308, 214–216. <https://doi.org/10.1126/science.1111656>
- Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of Mind is contagious: You catch it from your sibs. *Child Development*, 65(4), 1228–1238. <https://doi.org/10.2307/1131316>
- Peterson, C. C., & Siegal, M. (1999). Representing inner worlds: Theory of Mind in autistic, deaf, and normal hearing children. *Psychological Science*, 10(2), 126–129. <https://doi.org/10.1111/1467-9280.00119>
- Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic Theory of Mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, 26(9), 1353–1367. <https://doi.org/10.1177/0956797614558717>
- Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., ... Ruffman, T. (2018). Do infants understand false beliefs? We don't know yet – A commentary on

- Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development*, 48, 302–315. <https://doi.org/10.1016/j.cogdev.2018.09.005>
- Poulin-Dubois, D., & Yott, J. (2018). Probing the depth of infants' theory of mind: disunity in performance across paradigms. *Developmental Science*, 21(4), e12600. <https://doi.org/10.1111/desc.12600>
- Povinelli, D. J., & De Blois, S. T. (1992). Young children (*Homo sapiens*) understanding of knowledge formation in themselves and others. *Journal of Comparative Psychology*, 106(3), 228–238.
- Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, 46, 40–50. <https://doi.org/10.1016/j.cogdev.2017.10.004>
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, 4, 515–526. <https://doi.org/10.1016/j.celrep.2011.1011.1001.7>
- Priewasser, B., Rafetseder, E., Gargitter, C., & Perner, J. (2018). Helping as an early indicator of a theory of mind: Mentalism or Teleology? *Cognitive Development*, 46, 69–78. <https://doi.org/10.1016/j.cogdev.2017.08.002>
- Psouni, E., Falck, A., Boström, L., Persson, M., Sidén, L., & Wallin, M. (2018). Together I can! Joint attention boosts 3- to 4-year-olds' performance in a verbal false-belief test. *Child Development*. <https://doi.org/10.1111/cdev.13075>
- Rakoczy, H. (2010). Executive function and the development of belief-desire psychology. *Developmental Science*, 13(4), 648–661. <https://doi.org/10.1111/j.1467-7687.2009.00922.x>
- Rakoczy, H. (2012). Do infants have a Theory of Mind? *British Journal of Developmental Psychology*, 30(1), 59–74. <https://doi.org/10.1111/j.2044-835X.2011.02061.x>
- Rakoczy, H., Bergfeld, D., Schwarz, I., & Fiske, E. (2015). Explicit Theory of Mind is even more unified than previously assumed: Belief ascription and understanding aspectuality emerge together in development. *Child Development*, 86(2), 486–502. <https://doi.org/10.1111/cdev.12311>

- Rhodes, M., & Brandone, A. C. (2014). Three-year-olds' theories of mind in actions and words. *Frontiers in Psychology, 5*(263), 1–8. <https://doi.org/10.3389/fpsyg.2014.00263>
- Richardson, H., Koster-Hale, J., Caselli, N., Magid, R. W., Benedict, R., Olson, H., ... Saxe, R. (2018). How Language Facilitates Theory of Mind Development: Behavioral and FMRI Evidence from Individuals with Delayed Access to Language. *PsyArXiv Preprints*. <https://doi.org/10.31234/OSF.IO/8KRMB>
- Rubio-Fernández, P. (2018a). Publication standards in infancy research: Three ways to make Violation-of-Expectation studies more reliable. *Infant Behavior and Development, 181*. <https://doi.org/10.1016/j.infbeh.2018.09.009>
- Rubio-Fernández, P. (2018b). What do failed (and successful) replications with the Duplo task show? *Cognitive Development, 53*. <https://doi.org/10.1016/j.cogdev.2018.07.004>
- Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science, 24*(1), 27–33. <https://doi.org/10.1177/0956797612447819>
- Rubio-Fernández, P., & Geurts, B. (2016). Don't mention the marble! The role of attentional processes in false-belief tasks. *Review of Philosophy and Psychology, 7*(4), 835–850. <https://doi.org/10.1007/s13164-015-0290-z>
- Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and Theory-of-Mind understanding. *Child Development, 73*(3), 734–751. <https://doi.org/10.1111/1467-8624.00435>
- Sabbagh, M. A., Moses, L. J., & Shiverick, S. (2006). Executive functioning and preschoolers' understanding of false beliefs, false photographs, and false signs. *Child Development, 77*(4), 1034–1049. <https://doi.org/10.1111/j.1467-8624.2006.00917.x>
- Sabbagh, M. A., & Paulus, M. (2018). Replication studies of implicit false belief with infants and toddlers. *Cognitive Development, 46*, 1–3. <https://doi.org/10.1016/j.cogdev.2018.07.003>
- Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their Way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1255–1266. <https://doi.org/10.1037/a0018729>

- Santiesteban, I., Shah, P., White, S., Bird, G., & Heyes, C. (2015). Mentalizing or submentalizing in a communication task? Evidence from autism and a camera control. *Psychonomic Bulletin and Review*, *22*(3), 844–849. <https://doi.org/10.3758/s13423-014-0716-0>
- Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and Theory of Mind: A study of deaf children. *Child Development*, *78*(2), 376–396. <https://doi.org/10.1111/j.1467-8624.2007.01004.x>
- Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology*, *141*(3), 433–438. <https://doi.org/10.1037/a0025458>
- Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. *Cognition*, *162*, 27–31. <https://doi.org/10.1016/j.cognition.2017.01.018>
- Schuwerk, T., Priewasser, B., Sodian, B., & Perner, J. (2018). The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt. *Royal Society Open Science*, *5*(5), 172273. <https://doi.org/10.1098/rsos.172273>
- Scott, R. M. (2017). Surprise! 20-month-old infants understand the emotional consequences of false beliefs. *Cognition*, *159*, 33–47. <https://doi.org/10.1016/j.cognition.2016.11.005>
- Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249. <https://doi.org/10.1016/j.tics.2017.01.012>
- Scott, R. M., Baillargeon, R., Song, H. J., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology*, *61*(4), 366–395. <https://doi.org/10.1016/j.cogpsych.2010.09.001>
- Scott, R. M., He, Z., Baillargeon, R., & Cummins, D. (2012). False-belief understanding in 2.5-year-olds: Evidence from two novel verbal spontaneous-response tasks. *Developmental Science*, *15*(2), 181–193. <https://doi.org/10.1111/j.1467-7687.2011.01103.x>
- Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive Psychology*, *82*, 32–56. <https://doi.org/10.1016/j.cogpsych.2015.08.003>

- Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., ... Csibra, G. (2010). Absence of spontaneous action anticipation by false belief attribution in children with autism spectrum disorder. *Development and Psychopathology*, 22(2), 353–360. <https://doi.org/10.1017/S0954579410000106>
- Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science*, 22(7), 878–880. <https://doi.org/10.1177/0956797611411584>
- Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous Theory of Mind in Asperger syndrome. *Science*, 325(5942), 883–885. <https://doi.org/10.1126/science.1176170>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology*, 143(2), 534–547. <https://doi.org/10.1037/a0033242>
- Sodian, B., & Kristen-Antonow, S. (2015). Declarative joint attention as a foundation of theory of mind. *Developmental Psychology*, 51(9), 1190–1200. <https://doi.org/10.4161/epi.6.3.14196>
- Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, 13(6), 907–912. <https://doi.org/10.1111/j.1467-7687.2009.00946.x>
- Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18(7), 587–592. <https://doi.org/10.1111/j.1467-9280.2007.01944.x>
- Southgate, V., & Vennetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, 130, 1–10. <https://doi.org/10.1016/j.cognition.2013.08.008>
- Sprung, M., Perner, J., & Mitchell, P. (2007). Opacity and discourse referents: Object identity and object properties. *Mind and Language*, 22(3), 215–245. <https://doi.org/10.1111/j.1468-0017.2007.00307.x>

- Sullivan, K., & Winner, E. (1993). Three-year-olds' understanding of mental states: The influence of trickery. *Journal of Experimental Child Psychology*, *56*, 135–148.
- Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580–586. <https://doi.org/10.1111/j.1467-9280.2007.01943.x>
- Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology*, *30*(1), 30–44. <https://doi.org/10.1111/j.2044-835X.2011.02046.x>
- Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology*, *30*(1), 172–187. <https://doi.org/10.1111/j.2044-835X.2011.02067.x>
- Tobii AB. (2016). *Tobii Studio User's Manual v3.4.5*. Tobii AB.
- Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge: Harvard University Press.
- Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences*, *115*(34), 8491–8498. <https://doi.org/10.1073/pnas.1804761115>
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(05), 675–91; discussion 691–735. <https://doi.org/10.1017/S0140525X05000129>
- Tomasello, M., & Haberl, K. (2003). Understanding attention: 12- and 18-month-olds know what is new for other persons. *Developmental Psychology*, *39*(5), 906–912. <https://doi.org/10.1037/0012-1649.39.5.906>
- Tomasello, M., & Rakoczy, H. (2003). What makes human cognition unique? From individual to shared to collective intentionality. *Mind and Language*, *18*(2), 121–147. <https://doi.org/10.1111/1468-0017.00217>
- Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs? *Infancy*, *15*(4), 434–444. <https://doi.org/10.1111/j.1532->

7078.2009.00025.x

- van der Wel, R. P. R. D., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*(1), 128–133. <https://doi.org/10.1016/j.cognition.2013.10.004>
- Wang, L., & Leslie, A. M. (2016). Is implicit Theory of Mind the “real deal”? The own-belief/true-belief default in adults and young preschoolers. *Mind & Language*, *31*(2), 147–176. <https://doi.org/10.1111/mila.12099>
- Warneken, F., & Tomasello, M. (2007). Helping and cooperation at 14 months of age. *Infancy*, *11*(3), 271–294. <https://doi.org/10.1111/j.1532-7078.2007.tb00227.x>
- Wellman, H. M. (2014). *Making minds: How Theory of Mind develops*. Oxford: Oxford University Press.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of Theory-of-Mind development: The truth about false belief. *Child Development*, *72*(3), 655–684. <https://doi.org/10.1111/1467-8624.00304>
- Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind tasks. *Child Development*, *75*(2), 523–541. <https://doi.org/10.1111/j.1467-8624.2004.00691.x>
- Wellman, H. M., Lopez-Duran, S., LaBounty, J., & Hamilton, B. (2008). Infant attention to intentional action predicts preschool theory of mind. *Developmental Psychology*, *44*(2), 618–623.
- Wellman, H. M., Phillips, A. T., Dunphy-Lelii, S., & LaLonde, N. (2004). Infant social attention predicts preschool social cognition. *Developmental Science*, *7*(3), 283–288. <https://doi.org/10.1111/j.1467-7687.2004.00347.x>
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)
- Yeung, H. H., Denison, S., & Johnson, S. P. (2016). Infants' looking to surprising events: When eye-tracking reveals more than looking time. *Plos One*, *11*(12), e0164277. <https://doi.org/10.1371/journal.pone.0164277>

- Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief? *British Journal of Developmental Psychology*, *30*(1), 156–171. <https://doi.org/10.1111/j.2044-835X.2011.02060.x>
- Yott, J., & Poulin-Dubois, D. (2016). Are infants' Theory of Mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development*, *17*(5), 683–698. <https://doi.org/10.1080/15248372.2015.1086771>
- Zmyj, N., Prinz, W., & Daum, M. M. (2015). Eighteen-month-olds' memory interference and distraction in a modified A-not-B task is not associated with their anticipatory looking in a false-belief task. *Frontiers in Psychology*, *6*, 857. <https://doi.org/10.3389/fpsyg.2015.00857>