# The Development of
# Recursive Meta-Representational Theory of Mind

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsstudiengang Behavior and Cognition (BeCog)
der Georg-August University School of Science (GAUSS)

vorgelegt von
Lydia Paulin Schidelko
aus Recklinghausen

Göttingen, 2023

Betreuungsausschuss

Prof. Dr. Hannes Rakoczy
Kognitive Entwicklungspsychologie,
Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Prof. Dr. Julia Fischer
Deutsches Primatenzentrum
Kognitive Ethologie, Universität Göttingen

Prof. Dr. Markus Steinbach
Seminar für Deutsche Philologie, Universität Göttingen

Mitglieder der Prüfungskommission

Referent: Prof. Dr. Hannes Rakoczy
Kognitive Entwicklungspsychologie,
Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Koreferentin: Prof. Dr. Julia Fischer
Deutsches Primatenzentrum
Kognitive Ethologie, Universität Göttingen

Weitere Mitglieder der Prüfungskommission

Prof. Dr. Markus Steinbach
Seminar für Deutsche Philologie, Universität Göttingen

Prof. Dr. Michael Waldmann
Kognitionswissenschaft und Entscheidungspsychologie,
Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Dr. Claudia Fichtel
Deutsches Primatenzentrum
Verhaltensökologie und Soziobiologie, Universität Göttingen

Dr. Tanya Behne
Kognitive Entwicklungspsychologie,
Georg-Elias-Müller- Institut für Psychologie, Universität Göttingen

Tag der mündlichen Prüfung: 10.5.2023

## Preliminary Note

The present thesis is a publication-based (cumulative) dissertation. It is based on three original articles that have been published in international peer-reviewed journals, and one original research article that is currently in preparation:

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task. *Frontiers in Psychology*, 12, 4573.

Schidelko, L. P., Huemer, M., Schröder, L. M., Lueb, A. S., Perner, J., & Rakoczy, H. (2022). Why Do Children Who Solve FB Tasks Begin to Find True Belief Control Tasks Difficult? A Test of Pragmatic Performance Factors in ToM Tasks. *Frontiers in Psychology*, 12, 797246.

Schidelko, L. P., Proft, M., & Rakoczy, H. (2022). How do children overcome their pragmatic performance problems in the true belief task? The role of advanced pragmatics and higher-order theory of mind. *PLOS ONE*, 27.

Schidelko, L. P., Baumann, L., Proft, M., & Rakoczy, H. (2023). *Do Theory of Mind and Mental Time Travel abilities build on joint cognitive foundations? A Comparison of development of recursive Theory of Mind and Mental Time Travel* [Unpublished manuscript]. Department of Developmental Psychology. University of Göttingen.

In the present thesis, I relate the contents of the four articles to a broader theoretical context, summarize the main findings, and provide an extensive joint discussion. The original articles are reprinted in Appendices A, B, C, and D. I served as the first author in all the articles and was responsible for (a) developing the theory, (b) designing, conducting, and supervising the experiments, (c) analyzing and interpreting the data, and (d) writing up and publishing the manuscripts. My thesis supervisor, Hannes Rakoczy, as well as my co-authors, supported me with helpful advice and discussion throughout the process.

I hereby declare that all parts of this dissertation were written by myself. The assistance of third parties was only accepted if scientifically justifiable and acceptable in regard to the examination regulations. Additionally, all sources have been quoted.

**Thank you**

To **Hannes Rakoczy** for his supervision, cooperation, and trust, enabling me to pursue my scientific ideas.

To **Britta Schünemann** and **Marina Proft** for instilling within me enthusiasm for the field and supporting me since my bachelor thesis.

To **Julia Fischer** and **Markus Steinbach** for their valuable feedback during the thesis committee meetings and for their support.

To **Claudia Fichtel**, **Michael Waldmann,** and **Tanya Behne** for agreeing to be members of the examination board.

To **Micheal Huemer** and **Josef Perner** for the great collaboration.

To the whole Kindsköpfe lab, especially **Regina Zörner** and **Marlen Kaufmann** and all **student assistants** and **interns** for making these and more projects possible, especially to **Anna Sophie Lueb, Cathrin Degen, Christina Schneider, Dorothea van Dyk, Jana Rechenburg, Laura Bertram, Lara Salmon, Leonie Baumann, Lia Künnemann, Marea Latzel, Maren Stoll, Marlene Meyer, Sophia Hegemann,** and **Sophie Pötzke.**

To **Britta Schünemann, Feride Nur Haskaraca-Kizilay, Isa Garbisch, Lisa Wenzel, Natalie Bleijlevens, Ricarda Bothe, Rowan Titchener,** and **Saba Amirhaftehran** for the companionship on the Ph.D. journey, lively discussions, great feedback that benefitted my work, and the amazing time we had during all our reading groups, conference travels and coffee breaks.

To my **parents**, my **sister,** my **brother,** and my **friends,** especially to **Jusch** and **Mareike** for their incredible moral support. To **Mareike** and **Britta** for proofreading earlier drafts of this thesis and to **Mila** for reminding me to take breaks.

To **Gabriel**, who believed in me even when I couldn't believe in myself, for giving me the courage to start this Ph.D. and enabling me to keep going and finish this thesis.

Finally, to **all the children,** the great mental time travelers and belief reasoners, who participated in the studies. It has been a pleasure working with you!

# Table of Contents

# 0. Summaries

## 0.1. English Summary

Domain-general theories of Theory of Mind development assume that the ability to understand mental states (representations) as such is based on a fundamental understanding of representations as representations (meta-representations) (Perner, 1991). Accordingly, Theory of Mind develops in the course of a cognitive revolution to meta-representational thinking at the age of about four years (Perner, 1991; Wellman et al., 2001). In subsequent development, more complex forms of metarepresentational Theory of Mind develop by recursive embedding of multiple mental representations, which additionally requires the ability of recursive operations ("A thinks that B thinks that C thinks … that *p*"). However, little is known about the developmental trajectories of this prolonged development of complex, recursive Theory of Mind. Moreover, recent research has cast doubt on the assumption that meta-representational Theory of Mind develops by age four (*e.g.,* Fabricius et al., 2010; 2021).

Therefore, the projects of this dissertation aim to investigate the emergence and development of meta-representational ToM from a domain-general perspective by addressing four questions: (Ia) When does meta-representational Theory of Mind emerge, (Ib) How does recursive meta-representational Theory of Mind develop throughout childhood, and what are developmental relations of ToM to (IIa) a general understanding of representations and (IIb) to the ability of recursive thinking?

These research questions were addressed in three projects. Due to the restrictions of the COVID-19 pandemic, data collection for most of the studies of the dissertation had to be conducted remotely. Therefore, a preliminary study (*N*= *188,* 3-4 years) established the new method of online data collection as a valid alternative to classical in-person test settings.

The first main project addressed the emergence of Theory of Mind in comparison to children's general understanding of representations (*N* = 88, 3-6 years). The results showed parallel and correlated performance in meta-representational thought providing evidence for a unified emergence of meta-representational thought around age four.

Building on the results of a unified onset, the second project investigated the subsequent development of more complex forms of recursive meta-representational thinking. In three studies (*N* = 232), six-to ten-year-old children were tested in first-to fifth-order Theory of Mind tasks. The results revealed major developmental progress in higher-order Theory of Mind and substantial correlations to general recursive embedding abilities. This provides evidence for the relatedness of complex Theory of Mind and higher-order recursive thinking.

To shed light on further cognitive correlates or foundations, the third project compared this developmental progress in higher-order Theory of Mind with children's Mental Time Travel abilities. Mental Time Travel refers to the ability to reason about past and future events. It has been argued that sophisticated Mental Time Travel requires the representation of how past or future events relate to the present state and is thus based on some form of meta-representation. Children's first-, second-, and third-order Theory of Mind and Mental Time Traveling abilities were compared ($N$ = 120, 3-8 years). The results showed parallel, stepwise developmental progress from first to third-order Theory of Mind and first-to third-order Mental Time Travel. The developmental trajectories were, however, only partly associated.

In summary, the results show a consistent onset of meta-representational thought around age four and major developmental progress in higher-order Theory of Mind in children between six and ten years. Emergence and subsequent development in Theory of Mind was (partly) related to development in general recursive thinking abilities and meta-representational thought in other domains. These findings are compatible with domain-general theories of Theory of Mind development: To a certain extent, the emergence of first-order Theory of Mind and developmental progress in higher-order Theory of Mind base on shared abilities as meta-representational thought and recursive reasoning. This evidence contributes to a more comprehensive understanding of ontogenetic Theory of Mind emergence and subsequent development in Theory of Mind and related abilities.

## 0.2. Deutsche Zusammenfassung

Bereichsübergreifende Theorien zur Entwicklung von Theory of Mind nehmen an, dass die Fähigkeit, mentale Zustände (Repräsentationen) als solche zu verstehen auf einem grundlegenden Verständnis von Repräsentationen als Repräsentation (Metarepäsentationen) beruht. Demnach entwickelt sich Theory of Mind im Laufe einer kognitiven Revolution zu metarepräsentationalem Denken im Alter von etwa vier Jahren (Perner, 1991; Wellman et al., 2001). Im weiteren Verkauf der kindlichen Entwicklung entstehen komplexere Formen metarepräsentationaler Theory of Mind durch rekursive Einbettung mehrerer mentaler Repräsentationen, was zusätzlich die Fähigkeit rekursiven Denkens voraussetzt ("A denkt, dass B denkt, dass C denkt ... dass *p*"). Bisher ist nur wenig über die Entwicklungswege dieser späteren Entwicklung von komplexer, rekursiver Theory of Mind bekannt. Darüber hinaus haben neuere Forschungsergebnisse Zweifel an der Annahme aufkommen lassen, dass sich metarepräsentationale Theory of Mind im Alter von vier Jahren entwickelt (*e.g.*, Fabricius et al., 2010; 2021).

Die Projekte dieser Dissertation zielten daher darauf ab, die Entstehung und Entwicklung von metarepräsentationaler Theory of Mind aus einer domänenübergreifenden Perspektive zu untersuchen. Dazu wurden folgende Forschungsfragen adressiert: (Ia) Wann entwickeln Kinder meta-repräsentationale Theory of Mind und (Ib) Wie entwickeln sich spätere Formen komplexer, rekursiver Theory of Mind? (II) Gibt es Zusammenhänge in der Entstehung und weiteren Entwicklung von ToM und (IIa) einem allgemeinen Verständnis von Repräsentationen, sowie (IIb) allgemeinen Rekursionsfähigkeiten?

Diese Forschungsfragen wurden in drei Projekten untersucht. Aufgrund der Covid-19-Pandemie fand die Datenerhebung größtenteils im Rahmen von Online-Studien statt. Daher wurde eine Vorstudie (*N*= 188, 3-4 Jahre) durchgeführt, in welcher diese neue Methode der Online-Datenerhebung als zuverlässige Alternative zu klassischen Präsenz-Testsettings validiert wurde.

Das erste Hauptprojekt der Dissertation untersuchte die Entstehung von Theory of Mind im Vergleich zu einem allgemeinen Verständnis von Repräsentationen. Dazu wurde das Verständnis drei- bis sechsjähriger Kinder (*N* = 88) von mentalen und nicht-mentalen Metarepräsentationen verglichen. Kinder zeigten alterstypische und parallele Performanz in metarepräsentationalen Aufgaben und abweichende Performanz in parallelen Aufgaben, die in ihrer repräsentationalen Struktur abwichen. Diese Ergebnisse sprechen für einen einheitlichen Beginn des meta-repräsentationalen Denkens in verschiedenen Domänen im Alter von vier Jahren.

Aufbauend auf den Ergebnissen einer einheitlichen Entstehung metarepräsentationalen Denkens untersuchte das zweite Projekt die Entwicklung komplexerer Formen rekursiver metarepräsentationaler Theory of Mind höherer Ordnung. In drei Studien ($N$ = 232) wurden sechs- bis zehnjährige Kinder in rekursiven Theory of Mind-Aufgaben (erste bis fünfte Ordnung) und in ihren allgemeinen Rekursionsfähigkeiten getestet. Die Ergebnisse zeigten große Entwicklungsfortschritte in Theory of Mind höherer Ordnung. Die Entwicklung korrelierte zudem mit Fähigkeiten rekursiven Denkens in anderen Bereichen. Dies spricht für einen Zusammenhang Theory of Mind höherer Ordnung und allgemeinen Rekursionsfähigkeiten in der kognitiven Entwicklung.

Das dritte Projekt kombinierte die Ideen aus den ersten beiden Projekten und verglich die Entwicklung rekursiver Theory of Mind mit Entwicklungen im Bereich der Mentalen Zeitreisefähigkeiten. Als Mentale Zeitreise wird die Fähigkeit bezeichnet, Repräsentationen vergangener und zukünftiger Ereignisse und deren Bezug zur Gegenwart zu repräsentieren. Basierend auf der Annahme, dass diese Fähigkeit ebenfalls auf (rekursivem) meta-repräsentationalem Denken beruht, wurden in diesem Projekt drei- bis achtjährige Kinder ($N$ = 120) in ihren Theory of Mind- und Mentale Zeitreisefähigkeiten verschiedener Rekursionsebenen verglichen. Die Ergebnisse zeigten einen schrittweisen, partiell korrelierten Entwicklungsverlauf von rekursiver Theory of Mind und Mentaler Zeitreise.

Zusammenfassend zeigen die Ergebnisse einen einheitlichen Beginn metarepräsentationalen Denkens im Alter von vier Jahren, und parallele, teilweise verbundene Entwicklung in Rekursionsfähigkeiten und metarepräsentationalem Denken in anderen Bereichen. Diese Befunde sind kompatibel mit bereichsübergreifenden Theorien zur Theory of Mind-Entwicklung und tragen zu einem umfassenderen Verständnis der ontogenetischen Theory of Mind-Entstehung und weiterführenden Entwicklung und zugrundeliegenden Fähigkeiten bei.

# 1. General Introduction

> "[The] understanding of mental representation is fundamental to our folk psychology, to our theory of mind. It is so deeply rooted in our way of thinking that it is almost impossible for us to imagine what it would be like not to think like this. There was a time, however, when we didn't."
>
> (Janet Wilde Astington, 1993)

Social interactions play a significant role in our daily lives, as we communicate, cooperate, compete, or engage with other people in various ways. As we navigate these social interactions, we take into account our as well as others' beliefs, desires, and feelings. This requires an understanding of mental states (or: *mental representations*). This everyday psychology is also referred to as *folk psychology* or *Theory of Mind* (ToM). It is considered one of the most fundamental capacities to navigate a complex social world (Leslie et al., 2004). Because of its extensive relevance to our mature social life, ToM became a major subject of research in psychology and related fields over the last few decades.

One central debate in this field of research concerns the question of whether ToM relies on processes specialized to the purpose of ToM computations (domain-specific processes) or whether, and to what extent, ToM devotes to more general capacities that are shared with other cognitive abilities (domain-general processes).

One way to approach this debate is to consider the ontogenetic origins and trajectories of ToM. The questions that arise are, when and how do children acquire a ToM? To what extent is the emergence and subsequent developmental progress in ToM decoupled from or related to developmental progresses in other domains?

Answers to these questions contribute to a comprehensive understanding of the developmental trajectories and cognitive routes of ToM. These answers may shed light on the debate of domain-specificity versus domain-generality of this central human capacity.

In light of these questions, my dissertation aims to investigate how children's developing abilities in ToM relate to other cognitive capacities. In particular, it aims to present a general understanding of representations and the ability of recursive thinking as fundamental for the emergence and development of ToM. Parallel developmental trajectories and associated performance in ToM and other cognitive domains would provide evidence for shared underlying foundations.

To approach these questions of developmental associations, we first need to clarify the developmental trajectories of ToM and ways to validly test these abilities in development. Thus, the dissertation will start by addressing questions (Ia) regarding the emergence of ToM, and (Ib) regarding the subsequent development of more complex forms of ToM and ways to measure these in development. This will provide us with the basis to investigate how the emergence and subsequent development in ToM relate to other cognitive abilities. Domain-general views have claimed that ToM relies on and develops with a general understanding of representations. Based on this idea, it has been suggested that more complex, higher-order forms of ToM rely on recursively embedding such representations which requires recursive thinking abilities. Thus, the dissertation investigates the developmental relations of ToM to (IIa) a general understanding of representations and (IIb) the ability of recursive thinking across development.

The dissertation aims to investigate these questions in the following way[1]: I will start by introducing ToM and its conceptual core – an understanding of mental representations. To this end, I will explain how ToM can be validly tested in development and draw the standard picture of ToM emergence based on established empirical evidence. I will then present recent empirical findings that seem to question this standard picture of ToM. Different accounts will be presented that aim to explain these recent findings. Some of these accounts interpret the recent findings in a way that conflicts with the standard picture of ToM emergence. I will discuss why these accounts are unlikely in light of preliminary evidence and present an alternative approach that is based on pragmatic limitation claims and is compatible with the standard picture of ToM. I will then propose a stringent way to empirically test the central prediction made by this approach. This provides the basis to address the first aim of this dissertation: the investigation of (Ia) the emergence of ToM and valid ways to test emerging ToM.

I will then delve into introducing more complex forms of ToM – higher-order ToM – which follow more protracted developmental trajectories after the emergence of ToM. I will present first attempts to empirically test such complex forms of ToM over the course of childhood. This would provide a foundation to address the second aim of this dissertation: (Ib) to expand upon

[1] I will outline the structure of the theoretical and empirical background of the dissertation in a rather detailed way because the investigation of (Ia) the emergence of ToM requires an excursus on a specific set of recent findings and a review of different approaches to interpret these findings. The detailed outline ought to give an overview of the structure of the dissertation and, at the same time, provides as explanation for the raisons d'être of the different parts of this dissertations.

methods to test higher-order ToM, displaying the trajectory of protracted development of complex, higher-order forms of ToM.

Armed with ways to clarify (Ia) the emergence and (Ib) subsequent development in ToM, I will address the main questions of this dissertation, the relation of ToM to (IIa) a general understanding of representations and (IIb) to the ability of recursive thinking. To this end, I will introduce the idea that Mental Time Travel abilities rely on a similar understanding of representations. Additionally, I will argue that more complex forms of Mental Time Travel involve recursive thinking to embed these representations within each other.

Finally, I will present the rationale and central findings of the three main projects of this dissertation. In the General Discussion of this dissertation, I will interpret these findings in light of domain-general ideas of ToM and will propose directions for future research. In the first part of the General Discussion, I will focus on a general understanding of representations and discuss differences in pragmatic demands in tasks used to test this understanding in different domains. In the second part of the General Discussion, I will discuss the role of Executive Functions in ToM development and the potential joint underlying cognitive structures of Executive Functions, ToM development, and other representational abilities.

## 2. Theory of Mind

In our daily lives, we perceive ourselves and others as rational agents who perform rational actions that are guided by their mental states about the world. Fundamentally, rational behavior is guided by *cognitive* mental states, such as beliefs and *conative* mental states, such as desires: Rational agents perform actions based on how they believe the world is and how they desire it to be. From a first-person perspective, mental states provide rational agents with practical reasons to act. From a third-person perspective, this conceptual framework can be used to predict and explain rational behavior. This capacity to attribute mental states to others and to oneself and to predict and interpret people's behavior based on their mental states is referred to as *Theory of Mind* (Premack & Woodruff, 1978). Often, we attribute a combination of a belief and a desire in order to predict or explain rational actions (Davidson, 1963). For example, when we observe a person taking a flower vase and taking a big sip from it as if it was a refreshing drink, we might reason: "She took a sip from the flower vase [action] because she wanted to drink soda [desire] and thought this was soda [belief]" (*cf.* Rakoczy, 2017). This makes ToM a fundamental capacity for practically every facet of our social life: Whenever we interact with each other, we monitor how others view the world and reason about what they might know, and what they might want or intend to do (Leslie et al., 2004).

In the following sections, I will approach the question of the emergence of ToM. In order to do so, I will start by introducing the idea of the conceptual core of ToM – an understanding of mental representations in the next section (Section 2.1 Meta-representation in Theory of Mind). The conceptual core provides direct implications on how to measure ToM in development in systematic and valid ways. These typical ToM tasks will be presented in the subsequent section (Section 2.2 How to Measure Theory of Mind). Armed with valid measures, the subsequent section (Section 2.3. Ontogeny of Theory of Mind) will present empirical findings of such measurement in ontogenetic development.

## 2.1. Meta-Representation in Theory of Mind

To have a ToM, it is essential to comprehend the representational function of the mind: Rational agents possess mental representations of various aspects of the world, which aim to represent how the world is (beliefs) or how it should be (desires). These representations depict the world *as being a certain way*. Rather than having direct access to reality, people construct and construe the world in their minds. Understanding that mental representations are mere representations of reality becomes relevant in cases where something is represented differently from how it really is. In these cases of *misrepresentation,* the representation of reality does not match actual reality – the *sense* differs from the *referent* of the representation. The *referent* of a representation refers to what the representation is representing, while the *sense* describes the specific way in which things are represented by the representation (Perner, 1991). Something (*i.e.*, the referent) is represented as being different (sense) from how it really is (Leekam & Perner, 1991). Grasping this distinction between sense and referent of mental representations, therefore, serves as indicator for an understanding of representations *as* representations which is fundamental to ToM (Perner, 1991).

To illustrate the point, let's consider again the example of the agent holding the belief that the flower vase was a glass of soda. Her mental representation refers to the situation in the world, that is the flower vase (*i.e.*, the referent). However, the mental representation represents the object differently – as being a drinking glass with soda (*i.e.*, the sense). Now, to interpret the agent's behavior (drinking from the flower vase) from our third-order perspective, it makes limited sense to rely on reality (the object being a flower vase) as it seems unlikely for anyone to knowingly drink the old and dirty water from a flower vase. Therefore, we need to rely on the agent's mental representation of reality instead. Grasping this representational function of the mind, we understand the mind as active, constructive, and interpretative (Astington, 1993). Importantly, the alternative representation of reality must not be confused with reality. Rather, it needs to be represented in parallel to, yet isolated from, reality. This is necessary to avoid the representation of reality getting lost amidst the alternative representation. To represent reality and an alternative representation of reality simultaneously, the alternative representation of reality has to be understood *as* mere representation (Perner, 1991).

Grasping this notion of representation, we can represent the agent's representation as representation. Understanding a representation as representation is also referred to as holding a *meta-representation*. In a strong reading, *meta-representing* is defined as 'represent(ing) the representational relationship itself' (Pylyshyn, 1978, p. 593). The *representational relationship* links the representation with what the representation represents. This relationship between the representation and what it represents has to be mentally modeled. Hence, meta-representation makes ToM a recursive operation. Recursive operations enable us to embed elements of the same kind within each other (Pinker & Jackendoff, 2005). In the context of ToM, this means embedding representations within representations by representing the relation between the misrepresentation and the actual reality (Perner, 1991; Pylyshyn, 1978).

To summarize, the conceptual core of ToM is meta-representational thought (Perner, 1991). It is necessary to address the conceptual core of meta-representation in great detail as this concept provides the theoretical basis to address the question of how to empirically test for ToM. In the next section, I will deduce implications and criteria for measuring ToM from the definition of meta-representation. Additionally, I will present a standard test for ToM in children to further address the aim of investigating the emergence of ToM in development.

## 2.2. How to Measure Theory of Mind

The meta-representational signature of a full-fledged ToM offers stringent criteria for how to measure developing ToM skills in systematic and valid ways.

As concluded in the last section, ToM conceptually relies on meta-representational thought which is required when reasoning involves alternative representations of reality. Hence, testing ToM requires distinguishing between reasoning based on the actual state of the world and reasoning based on an alternative representation of the world (Dennett, 1978). These considerations led to the development of the classic ToM False Belief (FB) Task by Wimmer and Perner (1983). Participants tested in FB tasks are required to ascribe a false (outdated) belief to a story protagonist and predict her action based on this belief. In other words, the story protagonist's belief needs to be represented as an alternative representation of the world besides the participants' own representation of reality (Perner, 1991; Wimmer & Perner, 1983)

In the classic version of the task, known as the change-of-location or unexpected transfer task, participants are presented with a story where the protagonist, Maxi, puts his chocolate into a blue cupboard before leaving the scene. In Maxi's absence, his mother moves the chocolate from the blue cupboard to a green cupboard. Participants are then asked where Maxi will look for his chocolate upon his return. ("Where will Maxi look for his chocolate?") (Wimmer & Perner, 1983). To succeed in the task, participants have to predict that Maxi will look for the chocolate in the blue cupboard. They have to mentally represent Maxi's misrepresentation (i.e., the chocolate being in the blue cupboard) of the current situation in the world (i.e., the chocolate being in the green cupboard). This requires participants to understand that Maxi's behavior is based on his subjective mental representation of the world, which is different from objective reality (Wimmer & Perner, 1983).

An important aspect of the task is that the storyline leads to a clear dissociation between how Maxi represents the world and how the participants represent the world. This dissociation ensures that participants cannot base their answer on reality, but rather must infer the action based on the story protagonist's subjective belief. This structure is elegant in at least two ways. Firstly, the test question tests indirectly for participants' FB understanding by asking for a belief-based action prediction. Thus, the answer does not require explicitly verbalizing the mental state by using mental verbs or complementary sentence structures ("Maxi believes that …"). It merely requires pointing gestures to the location Maxi will go to (i.e., the green or blue cupboard). Second, failure in the change-of-location FB task is not random. Participants who cannot (yet) represent the story protagonist's misrepresentation and infer the action from this, predict behavior on the basis of current reality. This is a reasonably successful strategy: If Maxi wants to have the chocolate, he will (or should) go to where the chocolate is (Perner et al., 2018; Priewasser et al., 2013). The strategy actually works in most cases as beliefs aim to and usually do align with reality. However, it fails in cases where the regular relationship between rational, belief-based actions and reality is disrupted (Doherty, 2009; Perner, 1991; Wimmer & Perner, 1983).

Another classic FB task is the unexpected content task (Hogrefe et al., 1986; Perner et al., 1987). It follows a similar structure as the change-of-location task but includes participants themselves having to learn that they held a FB before they are asked to ascribe this FB to another person. In this task, participants are shown a familiar container, for example, a *Smarties®* box, and are asked what is inside. After the participants answer (usually "Smarties"), they are shown that the box actually contains pencils. Participants are then asked to remember what they themselves thought was inside the box when they first saw it (own-belief question). After that, they are asked what they think someone else, for example, their friend, will believe is inside the box when first shown the box (other-belief question). To succeed in the task, participants have to first remember their own former beliefs, i.e., the answer to the question of what is inside the box. Then, they need to mentally represent someone else's current misrepresentation (*i.e.*, the box containing Smarties) of the situation in the world (i.e., the box containing pencils) (Hogrefe et al., 1986; Perner et al., 1987).

Armed with signatures of meta-representational ToM and standard ways of measurement in FB tasks, I will present empirical evidence to give a broad overview of ontogenetic development from precursors in social cognition to emergence of meta-representational ToM in the next section.

## 2.3. Ontogeny of Theory of Mind

### 2.3.1. Precursors and Early Forms of Theory of Mind

Early sensitivity to social stimuli enables infants to form early manifestations of skills and concepts during their early years, later forming into a part of ToM. Even though these skills provide an impressively early understanding of many social situations, they show clear signature limitations when it comes to meta-representational thought.

In their early months, infants' sensitivity and preference for social stimuli are revealed in early forms of interactions. At around nine months, infants begin to understand others' perceptions and the goals of their actions. In dyadic interactions, infants engage with their environment and soon after, they engage with others and external objects (triadic interactions) (*e.g.*, Brooks & Meltzoff, 2002; Csibra et al., 1999; Gergely et al., 1995; Moll & Tomasello, 2004). This marks the onset of a perception-goal-psychology: infants and toddlers reason based on this simple form of ToM which allows them to track what others perceive and which goals they have. In the second and third years of life, children evolve a perception-goal psychology with an increasingly flexible and sophisticated understanding of how agents' goals and preferences can differ (Wellman & Liu, 2004). This involves attributions of simple mental states and teleological action explanation and prediction. On a conative level, they appreciate that different agents can have different aims or preferences (*e.g.*, Repacholi & Gopnik, 1997). On a cognitive level, children can track that different agents can perceive different things. For example, from their second year, children can understand whether an agent can or cannot see things they themselves can see, or vice versa. This is often referred to as *Level-I perspective-taking* (Flavell et al., 1978). Importantly, acknowledging the preferences and visual perspectives of different agents does not necessarily require the ability to reason about conflicting versions of reality and thus, does not reflect meta-representational thought. Rather, such differences can be easily identified to align (*e.g.*, “eating broccoli is good for agent A and eating crackers is good for agent B”; Rakoczy et al., 2007; or “Agent A sees object X and agent B sees object Y”; Flavell et al., 1981; Perner et al., 2003). While perception-goal psychology allows an accurate social understanding in a variety of such situations, specific cases clearly show the signature limits of this simple form of ToM. These are cases in which a situation is misrepresented or a situation is represented in different ways by different agents and therefore, require meta-representational thought. Full-fledged meta-representational ToM develops only later in the course of the so-called *four-year revolution*.

### 2.3.2. Four-Year Revolution: Meta-Representational Theory of Mind

For more than four decades of research, standard FB tasks to test meta-representational ToM were administered to children from broad age groups and various cultures. Results reveal a clear developmental pattern of performance. Children consistently succeed in tasks that tap a meta-representational grasp from around age four (Wellman et al., 2001; Wimmer & Perner, 1983). Tasks include the standard change-of-location task, the unexpected content task, and many variations of these tasks. Children come to solve these tasks around the same age and their performance is highly consistent and correlated across tasks, as well as being robust to variations and consistent across cultures (Callaghan et al., 2005; Wellman et al., 2001). Their performance is mainly unaffected by task factors like, (I) the type of task used (*e.g.*, change-of-location task or unexpected content task), the target of the test question (*e.g.*, the question about one’s own previous FB or someone else’s FB) and, (III) the type of test question asked (*e.g.*, what will someone do, think, know or say) (Wellman et al., 2001).

Before the age of four, children systematically fail all kinds of FB tasks. Younger children consistently answer that Maxi will look for his chocolate in the green cupboard – according to reality – in the change of location task. Similarly, young children answer that others will think that the Smarties box contains pencils, not Smarties (*e.g.*, Wellman, 2001).

Also, from around age four, children begin to solve other tasks that require them to understand *subjective representations*. While children have to understand how an agent *mis*represents reality in classic FB tasks, other tasks require an understanding of *divergent* representations. All these tasks that require children to understand *how* an agent subjectively represents a situation are often summarized under the term *Level-II perspective-taking* tasks (in contrast to *Level-I perspective-taking* which only involves understanding that an agent sees something or does not see something) (Flavell et al., 1981). In typical Level-II perspective-taking tasks involving *divergent* representations, the participant and another agent look at an object (*e.g.*, a digit) between them that looks differently (like a ”6” or a “9”) depending on the side it is viewed from. Participants have to understand that the given situation appears different from different viewpoints (Flavell et al., 1981; Surtees et al., 2012).

This unified success in different tasks around age four reflects a conceptual change from the absence to the presence of meta-representational ToM and is also refered to as *four-year revolution* (Perner, 1991).

#### 2.3.3. General Transition to Meta-Representational Thought

In the course of the *four-year revolution*, children also begin to solve a range of other tasks. For example, children come to master tasks that involve appearance-reality distinctions. These appearance-reality tasks require participants to acknowledge deceptive appearances such that an object can look like one thing but really is something else. In these tasks, participants are shown a deceptive object, such as a sponge that looks like a rock. When they are allowed to touch the object, they realize that it is, in fact, a sponge. In two questions, children are asked to contrast what an object is and what it visually appears to be (“What is this really, really? Is it really, really a rock or really, really a piece of sponge?” and “When you look at this with your eyes right now, does it look like a rock or does it look like a piece of sponge?”) (Flavell et al., 1983, p. 102). Children acknowledge that the object really is a sponge that only looks like a rock around the age of four. They appreciate that their visual perception can misrepresent reality. Younger children, in contrast, give the same answer to both questions, saying that the object is a sponge that looks like a sponge. They do not yet understand that appearance can specify an object as something else than reality does (Flavell et al., 1983).

Taken together, children understand different kinds of subjective mental representations at the same age. Evidence from direct comparisons of these tasks shows that children do not only begin to master all tasks around the same age. Performance in different tasks is also strongly positively correlated. This joint success suggests that children's understanding of the subjectivity of mental representation develops in a systematic and unified manner (for a comprehensive overview, *see* Rakoczy, 2022). Thus, the four-year revolution marks the emergence of the subjective notion of mental representations, and with this, the transition to a meta-representational conceptual capacity (Leekam & Perner, 1991; Perner & Roessler, 2012). With this newly acquired capacity, children understand themselves and others as rational agents with subjective perspectives on the world that can differ in content and might deviate from reality.

In a broader context, the emerging meta-representational capacity is not limited to mental representations. Children start to solve tasks that involve non-mental representations around the same age. Subjective representations can be not only mental but also, for example, pictorial. In the False Sign task, for example, children need to understand that an actual situation is different from how a signpost represents it. In this task, in analogy to the change-of-location FB task, a signpost in a story scenario indicates a state of affairs (*e.g.*, that an ice-cream van is in Location 1). The object then moves to a new location (Location 2), but the signpost is not changed accordingly and therefore becomes a false sign (Parkin, 1994). Children have to mentally represent the signpost's representation of the world (*i.e.*, the ice-cream van being at location 1) in addition to their own current representation of the world (i.e., the ice-cream van being at location 2). Just like Maxi misrepresents the situation in the FB task, the current reality is misrepresented by the signpost in the False Sign task. Irrespective of considerable differences in task structure and format, children succeed in these tasks requiring understanding of mental or external representational at the same time. Additionally, performance across tasks is highly correlated (Bowler et al., 2005; Iao & Leekam, 2014; Leekam et al., 2008; Sabbagh et al., 2006). This may indicate a deeper underlying cognitive transition: in the course of the four-year revolution, children acquire a broad meta-representational conception of all kinds of representations – including mental and non-mental representations.

What develops in preschool years and what might explain why children come to solve these different types of tasks in tandem, is an underlying general capacity to coordinate multiple subjective perspectives and to solve "perspective problems" (Perner et al., 2003; Perner & Roessler, 2012). Perspective problems can pertain to all kinds of representations – mental, pictorial, or linguistic. Understanding a given situation constitutes a perspective problem when two representational contents cannot simply be combined conjunctively (by "and") – without explicitly relativizing the contents to representational standpoints. For example, "the chocolate is in the green cupboard and the chocolate is in the blue cupboard" or "the ice-cream van is at the house and the ice-cream van is at the church" are incoherent. Only when they are relativized to standpoints, can the two contents be conjoined: "From Maxi's perspective, the chocolate is in the blue cupboard and it really is in the green cupboard"; "According to the signpost, the ice-cream van is at the house, and it really is at the church" (Fizke et al., 2014; Perner et al., 2003). The standard picture of ToM development is thus that children develop a fully-fledged ToM around the age of four in the course of a major developmental transition to meta-representational perspective understanding (Leekam et al., 2008; Perner, 1991; Wellman et al., 2001).

### 2.3.4. Summary: Standard Picture on Theory of Mind Development

The standard picture of ToM development is thus that children develop ToM around the age of four. In contrast, younger children reason about their social world with simpler perception-goal inferences. Evidence for the emergence of ToM around age four comes from studies that ask for judgments about mental states or belief-based actions and thus require the participant to represent *subjective* mental representations. The ability to understand subjective mental representations is typically assessed through tasks that involve recognizing when an agent misrepresents reality or when there are divergent representations of reality based on different perspectives (Flavell et al., 1981; Wellman et al., 2001; Wimmer & Perner, 1983). Success in such tasks is closely related to tasks requiring meta-representational thought in a non-mental context (Leekam et al., 2008). This has led to the widely held belief that there is a unified onset of meta-representational ToM during the four-year revolution (Leekam et al., 2008; Perner, 1991; Perner, Brandl, et al., 2003; Wellman et al., 2001).

## 2.4. Challenging the Standard Picture

This standard picture has recently come under attack by different claims. On one hand, it has been proposed that classic FB tasks structurally underestimate children's ToM abilities. According to such underestimation claims, children much younger than four years of age show sophisticated forms of ToM (Bloom & German, 2000; Carruthers, 2013; Leslie et al., 2004; Scott & Baillargeon, 2017). On the other hand, classic FB tasks have also been questioned to overestimate children's ToM abilities. Overestimation claims suggest that four-year-old children still rely on simpler reasoning strategies and develop true ToM only later in development. Both underestimation and overestimation claims need to be considered to approach the overarching question of the emergence and subsequent development of meta-representational thought. The next section (2.4.1 Underestimation claims: Implicit Theory of Mind) will shortly introduce the idea of more indirect measures of ToM that ought to detect earlier forms of ToM. However, the current empirical situation is highly complex and does not allow clear interpretations of earlier meta-representational ToM capacities based on these indirect measures. More stringent evidence is needed to interpret how far these results challenge the standard picture of ToM. Therefore, I will only shortly give an overview of the current state of evidence in light of potential underestimation claims.

The focus will lie on overestimation claims. This will require an excursus on a specific set of recent findings and a review of accounts that aim to explain these results. The accounts vary in whether they are compatible or incompatible with the standard picture of ToM emergence. Thus, it will be thus necessary to evaluate the different accounts in detail and work out empirical test cases to test them to address the first aim of this dissertation: the investigation of (Ia) the emergence of ToM and valid ways to test emerging ToM.

### 2.4.1. Underestimation Claims: Implicit Theory of Mind

The standard interpretation of ToM development is subject to one fundamental concern, namely that classic FB tasks may have underestimated earlier ToM competence due to extraneous performance factors associated with high linguistic and cognitive task demands (Bloom & German, 2000; Carruthers, 2013; Leslie et al., 2004; Scott & Baillargeon, 2017). Based on these considerations, more indirect measures of ToM with reduced verbal and other cognitive task demands have been developed. One group of tasks is based on eye movement as an indicator of underlying cognitive processes. The main idea is that the participants' looking pattern reflects their underlying belief tracking processes (*see e.g.*, Onishi & Baillargeon, 2005; Schneider et al., 2012; Southgate et al., 2007; Surian & Geraci, 2012). Other tasks use spontaneous behavior to conclude whether children take into account what another agent believes when interacting and communicating with them (e.g., Buttelmann et al., 2009).

Studies that have used these indirect ToM measures with infants have called into question standard interpretations of the four-year revolution. Evidence from studies using looking time and interactive paradigms have indicated that even one to two-year-old infants show sophisticated belief tracking capacities. For example, in a change-of-location story, one to two-year-old children have been found to anticipate an agent to approach the location where he falsely believes the object to be (*e.g.*, Buttelmann et al., 2009; Onishi & Baillargeon, 2005; Southgate et al., 2007; Surian & Geraci, 2012). However, as impressive as these results were, these original findings could hardly be replicated (*e.g.*, Dörrenberg et al., 2018; Grosse Wiesmann et al., 2017; Kampis et al., 2021; Kulke et al., 2019; Kulke & Rakoczy, 2018; Poulin-Dubois & Yott, 2018). This failed replication calls into question the validity and reliability of implicit ToM capacities. In summary, the current empirical situation is highly complex and allows different interpretations about early ToM development, most of which admit some kind of relevant developmental progress in the course of the four-year revolution.

### 2.4.2. Overestimation Claims

On the other hand, there is a counterargument that classic FB tasks may actually overestimate children's ToM abilities. Although FB tasks are commonly considered fairly conservative, competence limitation accounts propose that they overestimate children's real ToM competence given at age four. The main claim of such competence limitation accounts is that even when children pass explicit FB tasks, they do not yet have meta-representational ToM. The empirical basis of these accounts is the failure of preschoolers in what are called True Belief (TB) tasks. The recent findings of the TB will be reviewed in the following section, after which I will present competence limitation accounts of ToM as well as an alternative explanation for this empirical phenomenon that is consistent with the standard interpretation of the four-year revolution. This will provide the foundation for empirical investigations of this dissertation into the developmental onset and progression of meta-representational thought.

#### 2.4.2.1. The "True Belief" Case

In the history of ToM research, many FB studies have included additional TB conditions to test children. The TB condition usually serves as a baseline measure to rule out the possibility that children's failure in the experimental FB condition is due to the general narrative structure of the task. As a result, TB tasks were often administered to young children and only rarely to older children.

In the change-of-location paradigm, the TB condition is identical to the FB condition, except that the protagonist witnesses the object transfer[2], resulting in a veridical belief (or: knowledge) of the object's location after it has been moved from location 1 to location 2. To maintain the

[2] Importantly, this TB version has to be distinguished from another, rarely used, TB version in which the object is not moved from location 1 to location 2 but it is moved and put back in the original location (location 1) in the protagonist's absence (Fabricius et al., 2010; Exp. 1). In this version, the protagonist holds a justified true belief (not knowledge) about the object being in location 1. This resembles so-called "Gettier cases" which were introduced in epistemology in order to demonstrate the difference between actual knowledge and mere (justified) true beliefs (Gettier, 1963). For the argumentation here, however, the knowledge-True Belief version is of main interest and therefore subject of further elaboration.

TB condition's structural similarity to the FB condition, the protagonist exits and re-enters the scene with no changes occurring to the object in their absence, either before or after the location change. The children are then asked the same test question: "Where will the protagonist look for her object?" (Fabricius et al., 2010, Exp. 2; Oktay-Gür & Rakoczy, 2017; Rakoczy & Oktay-Gür, 2020). Unlike the FB condition, the correct response in the TB condition is that the protagonist will look at location 2. Cases in which representations align with reality cannot unambiguously demonstrate meta-representational reasoning (Dennett, 1978): Children can come to the correct response without any belief ascription, answering teleologically based on the reality that the protagonist will look for the object where it actually is.

Typically, children younger than four years who are unable to pass the FB condition can successfully complete the TB condition (Wellman et al., 2001). However, recent studies administering the TB condition to a broader age range of children have yielded puzzling findings. Surprisingly, older children have been found to fail the TB control condition. Specifically, children aged four and older, who have begun to solve the FB task, have been found to consistently fail the TB task, systematically answering the TB questions incorrectly (Fabricius et al., 2010, 2021; Hedger & Fabricius, 2011; Huemer et al., 2023; Oktay-Gür & Rakoczy, 2017). The initial pattern of young children passing the TB condition but failing the FB condition reverses around this age. Children aged four and older succeed in passing the FB condition but fail the TB condition. This developmental pattern of younger children passing the TB condition and failing the FB condition, and the opposite for older children, leads to strong negative correlations of individual performances on the two versions of the tasks. It is noteworthy that the failure in the TB task persists into late childhood, with children only succeeding again from ages seven to ten. At this stage in development, children can pass both the FB and TB tasks for the first time. On an individual level, the strong negative correlations between performance on the two tasks disappear and turn into positive correlations. Overall, the development of performance on the TB task follows a U-shaped trajectory, with high performance in young children around age three, decreasing around age four when children solve the FB task and then rising again in later childhood between the ages of seven to ten. In sharp contrast, performance on the FB task remains constantly high in children older than four (Oktay-Gür & Rakoczy, 2017).

Different claims have been made to explain this empirical phenomenon. Some accounts cast doubt on children's competencies in ToM at age four (Fabricius et al., 2010, 2021; Perner et al., 2015), while other accounts rely on ideas of performance rather than competence limitations (Oktay-Gür & Rakoczy, 2017; Rakoczy & Oktay-Gür, 2020). While the former questions the standard picture of meta-representational ToM abilities at age four, the latter accounts are compatible with the standard interpretation of ToM development based on evidence from the classic FB tasks. This makes the TB task an important test case to approach the question of the emergence and development of meta-representational ToM. If the TB error reflects actual competence limitation in ToM after age four, this conflicts with the standard picture of ToM emergence. Thus, we need to evaluate these accounts to investigate (Ia) the emergence of ToM and valid ways to test emerging ToM.

#### 2.4.2.2. Competence Limitation Account I: Perceptual Access Reasoning

Some accounts claim that classic FB tasks, such as the change-of-location task and the unexpected content task, can be solved by using simpler strategies rather than by using proper meta-representational mental state reasoning. Children's failure in the TB tasks, however, uncovers these simpler strategies.

One such account suggested that children can solve FB tasks by using *Perceptual Access Reasoning* (Fabricius & Imbens-Bailey, 2000; Fabricius & Khalil, 2003). By applying two simple rules, children can end up with the correct response in these tasks without actually reasoning about their beliefs. The first rule is that seeing leads to knowing (*see* e.g., Perner, 1991; Wimmer et al., 1988). A person who sees a state of affairs knows that state of affairs. Conversely, a person who does not see a state of affairs does not know this state of affairs. The second rule is that knowing leads to correct behavior and not-knowing leads to incorrect behavior (Ruffman, 1996). A person who knows a state of affairs will act correctly – accordingly to this state of affairs. A person who does not know will act incorrectly (Fabricius & Khalil, 2003).

Let us first consider what these rules mean for success in FB tasks. Applying these two rules to FB tasks leads to the correct response. For example, in the change-of-location task, the agent leaves the scene and thus has limited perceptual access. Following the Perceptual Access Reasoning (PAR) account, "not seeing" leads to "not knowing" which, in turn, leads to "doing something wrong." As the agent in the change-of-location task wants to have her object, "incorrect" behavior would mean that she looks for the object where it is not. As the answer format of this task implies only two options for action (either looking for the object in location 1 or location 2), the only incorrect behavior predicted by PAR is to look in location 1. The same logic can be applied to the unexpected-content task. In contrast, using these rules in the TB task can lead to failure. Even though nothing happens to the object when the agent leaves the scene in the TB task, her perceptual access to the scene is interrupted. Expecting agents to act incorrectly whenever their perceptual access was limited or interrupted means to expect them to act wrong in the TB task.

Based on this analysis, the PAR claims that even four to five-year-olds are unlikely to have grasped the meta-representational nature of mental states (*see* Fabricius & Khalil, 2003). Instead, the development of ToM is claimed to be more prolonged than the four-year-revolution suggests (Fabricius & Imbens-Bailey, 2000). According to this account, children's reasoning strategies in ToM tasks undergo three stages. In the first stage, before age four, children reason based on reality. For example, they expect agents to search for their objects where they actually are. Children younger than four, thus, solve TB tasks but fail FB tasks. In the second stage, with onset around age four, children use PAR. They expect agents to act correctly when they have full perceptual access to the events but incorrectly if their access is limited. Children from age four expect the protagonist to act incorrectly in both the TB and FB task because the protagonist's perceptual access to the scene is interrupted in both tasks. However, this leads to success in the FB but failure in TB task. Only in stage three, children from around age six use competent belief reasoning and, therefore, master both TB and FB tasks (Fabricius et al., 2010, 2021; Fabricius & Imbens-Bailey, 2000; Fabricius & Khalil, 2003; Hedger & Fabricius, 2011).

The developmental sequence of the three stages – reality reasoning, PAR, and belief reasoning – would lead to success in the FB task from around age four and to a U-shaped performance pattern in the TB task.

Critically, these predictions of PAR only apply to a special variant of the change-of-location TB condition: the variant in which the protagonist leaves the scene and returns after the location change ("leave and return after location change"-TB condition used in Fabricius et al., 2010, *Exp. 2*; Fabricius et al., 2021, *Exp. Section VII*; Friedman et al., 2003). Only then, the protagonist's perceptual access is interrupted and the protagonist is perceived as being in a new situation without full perceptual access after her return[3]. In contrast, two other variants of the change-of-location TB condition do not predict the TB failure based on PAR. In the variant where the location change occurs after the protagonist's return, the protagonist is perceived as being in a new situation after the return but with full perceptual access to the location change ("leave and return before location change"-TB condition used in Oktay-Gür & Rakoczy, 2017, Exp. 1&2; Rakoczy & Oktay-Gür, 2020, Exp. 2 & 4). According to the PAR account, the protagonist is expected to act correctly in this condition. If the protagonist leaves the scene after

[3] The prediction of PAR account would also apply to justified True Belief (Gettier-case like) conditions (*see* Footnote 1; Gettier, 1963) as actual location change in happening during the protagonist's absence.

the transfer but does not come back before the test question is asked, it is less likely that children perceive the protagonist as being in a new situation and, hence, still attribute perceptual access ("leave but no return after location change"-condition used in Friedman et al., 2003, Exp. 2). Again, the protagonist is, according to the PAR account, expected to act correctly in this condition (Fabricius et al., 2010; Hedger & Fabricius, 2011; Huemer et al., 2023).

However, evidence of TB failure comes from all three variants of the TB condition (Fabricius et al., 2010; Huemer et al., 2023; Oktay-Gür & Rakoczy, 2017; Rakoczy & Oktay-Gür, 2020). This makes this PAR after age four and as an explanation for TB failure rather unlikely. Additionally, evidence from adaptions of the FB task provides further evidence against PAR after age four.

Adaptations of the classic version of the FB task, which controlled for mere ignorance-based rule reasoning, rejected concerns for PAR in these tasks. For example, in a study that introduced a third location (a third box) in the classic change-of-location task, the third location constituted a second empty and therefore also incorrect location. Applying PAR, the agent is only expected to act incorrectly (not searching at the correct location), but it is not specified at which wrong location. Four to five-year-old children did not randomly choose between the two wrong locations but consistently chose the object's initial location (location 1), indicating that children understand more about the protagonist's FB than just the protagonist's ignorance (Friedman et al., 2003; Perner & Horn, 2003).

In summary, proponents of the PAR account adopt the combination of children's success in the FB tasks and their failure in the TB condition at age four as evidence of the overestimation of children's ToM abilities at that age (Fabricius et al., 2010, 2021; Fabricius & Khalil, 2003; Hedger & Fabricius, 2011). However, evidence of success in modified FB tasks speaks against PAR around age four. Furthermore, evidence that children show difficulties not only in the versions of the TB task in which the agent lacks perceptual access but in various versions of the TB task contradicts the predictions of the PAR account.

#### 2.4.2.3. Competence Limitation Account II: Mental File Cards

Another competence limitation account predicts a similar U-shaped performance curve in the TB task. The Mental File Card Theory (MFCT) by Perner and colleagues (Perner et al., 2015) is a cognitive theory of mental representation underpinning ToM reasoning. It offers a framework for how information about objects and entities in the environment is stored, processed, and connected.

The theory builds on the basic idea that entities of the real world are represented in mental files. Mental files are representational structures that individuate their referent with a specific label (*e.g.*, “the ball”) and store predicative information about this referent (*e.g.*, “is red”). Whenever a new object is encountered in discourse or thought, a new mental file is created and anchored to the entity it represents. Predicative information about entities from one’s own perspective is stored in so-called *regular files,* while information other agents hold about the same entity is stored in *vicarious files* that are indexed to the respective agent. The content of regular files and vicarious files of the same entity may vary depending on what the other agent knows and thinks about the entity. A regular file is linked vertically to the corresponding vicarious file representing another agent’s perspective on the same entity. The vertical link allows for the switching of perspectives on the entity (from the own perspective to the agent’s perspective on the entity and vice versa). At the same time, the information contained in the regular file and the corresponding vicarious file are kept separate: predicative information from the regular file is not copied into the vicarious file if the other agent does not have access to this information (Perner et al., 2015).

An essential claim of the MFCT is that each file can only contain one individuating term (*i.e.*, one label). If an entity, such as a ball that rattles, is associated with a second label, for example, "the rattle," a new file is created. This new "rattle-file" is anchored to the same entity as the original "ball-file”. Predicative information about the entity that is captured under the new label (*i.e.,* “rattle”) is stored in the new file. Files that are anchored to the same entity are called *co-referential files* and can be linked horizontally. The horizontal link represents that co-referential files are anchored to the same referent and enable a free flow of information between mental files, making predicative information from both files available (Huemer et al., 2018; Perner et al., 2015; Perner & Leahy, 2016).

Throughout development, the architecture of mental files undergoes changes that help explain progress in ToM and related phenomena. Around the age of four, children learn to form horizontal links between mental files, connecting two files that represent the same referent. This enables them to comprehend identity statements like "The ball is also a rattle" (Perner et al., 2011). With these horizontal links allowing a flow of predicative information, children can combine the information that they have received under different labels. Around the same age, children begin to use vicarious files and form vertical links between regular files and vicarious files. For instance, when a child observes a protagonist putting an object into box 1, they create a regular file and a vicarious file for the object with the predicative information "located in box 1." If the object is then moved to box 2 without the protagonist witnessing the transfer, the predicative information is updated in the regular file of the object, but not in the vicarious file. The vicarious file that is vertically linked to the regular file of the object allows children to represent the agent's misrepresentation of the object's location. This enables them to solve classic FB tasks. The emergence of the ability to use vicarious files and form vertical links around age four explains why children begin to master explicit FB tasks at this time (Perner et al., 2015).

At the same time, the MFCT postulates that children's file management is still limited at age four. While children are generally able to deploy vicarious files indexed to other agents, they are still unable to prevent the deployment of unnecessary vicarious files. Furthermore, they are not yet able to coordinate both kinds of linking – horizontal and vertical – at the same time. According to MFCT, children can deploy only the vicarious files that capture the perspectives available to the indexed agent and coordinate both types of linking at the age of six (Huemer et al., 2018; Perner & Leahy, 2016).

This developmental trajectory of file architecture can explain why children between four and six tend to fail in TB tasks but pass FB tasks. More specifically, it can shed light on children's difficulty in a sub-group of TB tasks known as aspectual TB tasks. These tasks are the control version of aspectual FB tasks, which are based on the subjective notion of mental states that represent reality under certain aspects or descriptions (Searle, 1983).

In a recent version of an aspectual FB task (Rakoczy et al., 2015), children were presented with a dual-identity object, such as a pen that also doubles as a rattle, before a puppet entered the scene. The object was first presented as a pen in location 1 and then transferred as a rattle from location 1 to location 2. Although the child was aware of the object's dual identity, the puppet was not, and thus only saw a pen in location 1 and later heard a rattle in the experimenter's hand moving to location 2, without realizing that it was the same object (Rakoczy et al., 2015). In the parallel aspectual TB version, the protagonist is introduced to the dual-identity object, the pen that is also a rattle, and thus knows that the object is transferred to location 2. In the FB version of the aspectual task, children aged four and above successfully predicted that the puppet would look for the pen in location one because she did not know that the rattle that was transferred to location two was the same object. However, in the aspectual TB version, four- to six-year-old children failed: They answered that the puppet, as in the aspectual FB version, will look for the object in location 1 even though the puppet was informed about the dual identity of the object (Oktay-Gür & Rakoczy, 2017).

The MFCT provides the following explanation for these findings in the aspectual FB and TB tasks: Children below the age of four can solve the aspectual TB task correctly as they do not form a second file for the rattle but instead include the information that the ball rattles as predicative information in the regular pen-file. When the object is transferred, they add this location change as predicative information in the file. As they are not yet able to deploy any vicarious files, they answer the aspectual TB test question based on their regular file. As the file contains the predicative information about the object's location, children below the age of four answer the test question correctly. However, children in the second developmental phase can form two regular files for the pen and the rattle and link those files horizontally when they learn that both objects are the same. They are also able to deploy vicarious files for the pen and the rattle indexed to the protagonist and link the vicarious files vertically to their own regular files. However, they are not yet able to copy the horizontal link between the regular files and the vicarious files. When the object is then transferred as a rattle to location 2, this is only updated in the vicarious rattle-file but due to the missing horizontal link to the vicarious pen-file, the information update does not flow into the rattle-file. When being asked "where will the protagonist look for the pen," the child will answer incorrectly because her vicarious pen-file indexed to the protagonist still says ''is in location 1." At the same time, however, this leads to correct answers in the aspectual FB task. In this task, the protagonist does not know about the dual identity. The two unlinked vicarious files are in alignment with reality. Finally, at the third stage, from around the age of six, children become able to coordinate both horizontal and vertical linking at the same time, and thus master aspectual TB tasks. The developmental sequence of the three phases leads to mastery in the aspectual FB task from around the age of four and a U-shaped performance pattern in the aspectual TB task, comparable to the TB performance in standard TB versions (*e.g.*, Perner et al., 2015).

A recent set of studies involving FB and TB versions of aspectual belief tasks have empirically aligned with these explanations of the MFCT. The results showed that three-year-olds tended to pass TB and fail FB, while four to six-year-olds tended to pass FB and fail TB. Only from age six did children reliably solve both FB and TB tasks (Perner et al., 2015; Perner & Leahy, 2016). The authors concluded that children around the age of four, despite their mastery of the aspectual FB version, are not yet able to represent beliefs about identity. Thus, the MFCT suggests that children's ToM competencies around age four are still limited. Although children proficiently attribute misrepresentations and solve standard FB tasks, they still do not grasp the aspectuality of mental states. This calls into question the claim of a full-blown ToM around age four (Perner et al., 2015; Perner & Leahy, 2016).

In summary, proponents of the MFCT adopt the combination of children's success in aspectual FB tasks and their failure in the aspectual TB condition at age four as evidence of the overestimation of children's ToM abilities at that age (Perner et al., 2015). However, the MCFT's explanation of TB failure is limited to a certain subset of TB tasks, namely, aspectual TB tasks. Evidence shows that children do not only experience difficulties in these versions of the TB tasks but also in a range of different versions (*e.g.*, Oktay-Gür & Rakoczy, 2017), contradicting the predictions of the MFCT. The MFCT does not predict failure in the classic version in which a simple object (without any dual identity) is visibly moved to the new box witnessed by the protagonist. Based on performance according to MFCT in the second phase of development, children from age four are expected to solve this task without needing to coordinate vertical and horizontal links. Instead, they are expected to do so by deploying a vicarious file indexed to the protagonist that is linked vertically to their regular file (Perner et al., 2015; Perner & Leahy, 2016).

#### 2.4.2.4. Summary of Competence Limitation Accounts

Taken together, these findings of children's failure in TB tasks from age four have challenged the view of the full-blown explicit ToM in the course of the four-year revolution. The MFCT by Perner and colleagues (Perner et al., 2015) questions whether the belief reasoning competence in children around age four is full-blown. Instead, they argue that children's failure in the aspectual TB task reflects children's file architecture which allows them to reason about misrepresentations but does not allow them to reason proficiently about the aspectuality of beliefs until age six. In contrast, the PAR account of TB failure by Fabricius and colleagues (*e.g.*, Fabricius et al., 2010; 2021) calls into question whether children around age four engage in any ToM reasoning processes at all. They argue that children rely on heuristics of perceptual access. Even though both theories explain the TB performance pattern and have found empirical support for their predictions, the scope of application of both theories is restricted. The PAR account can only explain the U-shaped performance pattern in the standard change-of-location TB task in which the protagonist has a lack of perceptual access (*i.e.*, when she leaves the scene after the location change). In contrast, the MFCT only applies to TB tasks involving aspectuality. Thus, the two competence limitation accounts only offer selective explanations of children's U-shaped performance curves. None of the theories predict a U-shaped performance curve in a standard TB task, in which the protagonist leaves and returns before the location change. Empirically, the U-shaped performance curve occurs in a broad scope of TB versions including these standard TB tasks in which the protagonist leaves after the location change

(Oktay-Gür & Rakoczy, 2017). This can neither be explained by the PAR account nor by the MFCT.

An alternative approach is that children's difficulty with TB tasks reflects performance rather than real competence limitations in ToM. TB tasks might be artificially difficult because of pragmatic task demands. This idea of pragmatic performance limitation in the TB task will be presented in more detail in the following section. To foreshadow, accounts based on pragmatic performance limitation can, in contrast to the PAR account and the MFCT, explain TB failure in a broad scope of TB versions. While performance accounts question the standard picture of ToM emergence, alternative approaches of performance limitations do not question children's ToM emergence around age four (Rakoczy & Oktay-Gür, 2020).

The different conclusions about the emergence of ToM made by the accounts (PAR, MFCT, and pragmatics accounts) based on the results of the TB task make the TB task-relevant when addressing the question of ToM emergence. To approach the first overarching aim of this dissertation, investigating the emergence of ToM and finding valid ways to test for it, we need to answer the question of why children fail in the TB task.

#### 2.4.2.5. Performance Limitation Accounts: Pragmatics

In this section, I will introduce the idea that children fail the TB task because of its pragmatic task demands. I will start by introducing some general considerations about pragmatics in the TB task. Based on this, I will present a test case to investigate the hypothesis of pragmatics task demands in the TB task to determine, more specifically, the source of children's failure in the TB task in order to use this information to draw conclusions about children's ToM emergence more broadly.

Based on pragmatic task analysis, it was theorized that the TB task may be artificially difficult because it combines three factors that make the questions pragmatically confusing and demanding (Oktay-Gür & Rakoczy, 2017; Rakoczy & Oktay-Gür, 2020). (I) The TB test question is an academic test question. The child, the experimenter, and the story protagonist all see that the object is transferred from location 1 to location 2. Now, the experimenter asks the child where the protagonist believes that the object is. Regular questions are posed when a speaker asks for some information that she herself does not know (Searle, 1969). As the experimenter in the TB tasks knows the answer to the test question, she does not ask the question to receive new information from the child, but rather to test whether the child knows the answer too. In contrast to regular questions, academic questions have, thus, a more complex intentional and pragmatic structure: The speaker wants to know whether the interlocutor knows the answer that the speaker knows. For young children, this special question format seems especially difficult to understand (Siegal, 1999). (II) The task is highly trivial. In the TB storyline, an object is moved from one to another location and the protagonist observes that. The test question then is where the protagonist will look for that object. The answer is so obvious and common knowledge that children – even if pragmatically aware of academic questions – might find it difficult to make sense of the corresponding question and wonder why the experimenter asks such an easy question. (III) This may be particularly pronounced in the TB task where the test question points to the protagonist's subjective mental representation even though the protagonist shares common ground and is not subject to any error. Normally, this kind of mental state discourse about subjective mental representations has to have a certain point, such as someone's misrepresentation of the situation. Yet, the storyline of the TB task does not provide any possibility for error or misrepresentation: the child, the experimenter, and the story protagonist all see where the object is put and all see that the others observe that too. Such questions in the context of common knowledge without the possibility of error or misrepresentation are pragmatically unnatural (Papafragou et al., 2007). Therefore, the test question "What does she believe where the object is?" or "Where will she look for the object?" implies that there is an alternative perspective or misrepresentation involved and children may thus think that they must have missed something and look for a possible alternative perspective in the scenario (Rakoczy & Oktay-Gür, 2020).

In summary, the pragmatics analysis offers a performance-based explanation for children's failure in the TB task and can explain TB failure as a general phenomenon that applies to a broad scope of TB tasks. In contrast to the two competence limitation accounts introduced above, pragmatics accounts do not question general ToM abilities and are thus compatible with the claim of a four-year revolution. Quite the opposite: the analysis implies that the sensitivity to the crucial performance factors that make the task demanding is actually associated with ToM. Broadly speaking, ToM capacities lay the ground for developing a pragmatic understanding (*e.g.*, Happé, 1994). But when children start to develop ToM capacities and thus, a pragmatic understanding, both are still fragile. Their fragile pragmatics then leads them astray in the TB test. Young children, in contrast, without a sophisticated understanding of ToM are limited in their pragmatic language understanding and mostly use language literally (Matthews, 2014; but see Pouscoulous & Tomasello, 2020). According to pragmatics accounts, children in this stage of development thus should have no problems with the TB task. These considerations lead to the prediction of a negative association between ToM capacities (*i.e.*, the FB task) and the performance in the TB task: younger children are expected to fail the FB task but solve the TB task, but as soon as children come to master the FB task, they become confused by the TB task and thus, fail to answer the TB test question correctly (Oktay-Gür & Rakoczy, 2017; Rakoczy & Oktay-Gür, 2020).

#### 2.4.2.6. Competence vs. Performance Accounts: Test Cases

Competence and performance limitation accounts on TB performance differ in their predictions about the two aspects of TB tasks. First, the accounts make different predictions in which kind of TB tasks the U-shaped performance curve occurs. Second, manipulations of the extraneous task factors worked out in the pragmatic task analysis should affect TB performance according to pragmatics accounts but should have no impact on TB performance according to the two competence-based accounts. Once these critical pragmatic task factors have been removed or modified, children should answer the TB task correctly.

The first studies addressed both, (1) the kinds of TB tasks in which the U-shaped performance curve occurs and, (2) the effects of the manipulation of pragmatics task factors on TB performance. Regarding (1), children from age four showed difficulties in TB tasks of a broad variety. These were not limited to aspectual TB tasks or TB tasks in which the protagonist's perceptual access is interrupted just before the test question is asked. The evidence clearly speaks in favor of the pragmatics accounts, contradicting the predictions of both competence-based accounts. With regard to (2), the first studies have found preliminary evidence for the importance of the first two factors: (I) academic test questions in a (II) highly trivial task (Rakoczy & Oktay-Gür, 2020). Addressing both factor I and II, children were tested in a non-verbal version of the TB task (based on Call & Tomasello, 1999) that removed any (academic and trivial) questions. Children's performance was equally proficient in all age groups between four and seven, showing no evidence of a TB deficit (Rakoczy & Oktay-Gür, 2020).

But what about factor III: the question pertains to a rational agent's subjective perspective. Preliminary evidence for the influence of factor III on children's failure in the TB task comes from one recent study that compared FB/TB tasks with an analogous task that involved non-mental representations, the *False Photo task* (*based on* Zaitchik, 1990). The False Photo task was conceptualized as a non-mental analogue of the FB task (Zaitchik, 1990). In the False Photo task, the series of events is just like in an FB task (the object is put into location 1 and then moved from there to location 2). But rather than a human observer, a polaroid camera is in the scene and takes a photo of the object in location 1 before it is moved to location 2. While the photo develops, the object is then moved to location 2, and children are asked where the object is in the outdated ("false") photo. The results from several studies revealed similar developmental patterns in False Photo and FB tasks. The majority of three-year-olds failed in both tasks, while the majority of four-year-olds and older children passed both tasks (with slightly higher performance in the FB task) (Leekam & Perner, 1991; Zaitchik, 1990). Preliminary, that was interpreted in that the False Photo task is a test for meta-representational thought about non-mental representations (photos) (Sabbagh et al., 2006; Oktay-Gür & Rakoczy, 2017).

Based on this idea, a recent study used the False Photo task as a non-mental analogue to the FB tasks in order to explore the role that asking about subjective mental representations may play in making the TB task pragmatically complex. Four- to six-year-old children's performance in FB/TB tasks was therefore compared with their performance in analogous False/True Photo tasks. In the True Photo condition – in close analogy to the TB story – the camera took the photo when the object was already moved to the new location. The True Photo condition implements a crucial contrast case: a trivial academic test questions about a non-mental representation (of the photo). That constitutes a task that holds the first two factors – trivial, academic test question – constant while manipulating the third factor (asking about subjective mental representations). Results revealed that – consistent with previous findings – children's performance was comparable in the "false" conditions (FB and False Photo) such that children at ages four to six succeeded in both. But the performance was markedly different in the true conditions: while children, again, had difficulty in the TB tasks (performing below chance), they found the TP task easy (performed above chance) (Rakoczy & Oktay-Gür, 2020, *Exp. 2*). These findings thus provide prima facie evidence that it does matter whether academic trivial test questions pertain to subjective mental representations of rational agents (rather than to non-mental representations more generally).

However, the specific task used in that study – the False Photo task – makes the contrast between the FB and False Photo tasks difficult to interpret. The reason is that the False Photo task, strictly, speaking, does not involve a *mis*representation. From a theoretical point of view, the "false" photo is only *outdated* but not actually false (Perner & Leekam, 2008). Empirically, this concern is supported by findings that FB and False Photo, despite a similar age of onset, dissociate in typical and atypical development: There are no correlations between the task in typically developing children and autistic children have no difficulty with the False Photo task although they find FB tasks difficult (Davis & Pratt, 1995; Leslie & Thaiss, 1992; Perner & Leekam, 2008; Peterson & Siegal, 1998).

A better implementation of a task that involves perspective problems created by non-mental representation is the False Sign task (*see* also section 2.3.3. General Transition to Meta-Representational Thought). In the False Sign task, a signpost in a story scenario indicates a state of affairs (*e.g.*, that an object is in location 1) that misrepresents reality (the object is actually in location 2). From a theoretical point of view, the sign is not just outdated (like the photo showing the object in location is merely outdated but not false); it is misleading and false. Empirically, this analysis is corroborated by studies that show that False Sign and FB tasks are related beyond the same age of onset: performance in these tasks is highly correlated with typical and atypical development (and share analogous neural substrates and correlations with executive function; *see* e.g., Iao & Leekam, 2014; Leekam et al., 2008; Sabbagh et al., 2006).

In summary, pragmatic limitation accounts offer an explanation for the TB performance that is compatible with both children's failure in the different versions of the TB task and the onset of the meta-representational ToM around the same age. However, stringent evidence for the importance of all the contributing pragmatic factors of the task is still missing. Also, the developmental processes at the end of the U-shaped performance curve still raise the question of how children overcome their pragmatic difficulties in the TB task later in development.

So far, it is still an open question how children overcome their performance limitations in the TB task. Applying the idea of pragmatic confusion to the end of the U-shaped curve, the pragmatic analysis predicts that children recover and come to pass the TB task once they have made further progress in their pragmatic abilities (as adults show no difficulties in answering the TB test question correctly; Oktay-Gür & Rakoczy, 2017).

In view of these considerations, it can be hypothesized that increasing performance at the end of the U-shaped curve results from further developments in children's growing pragmatic abilities. Advances in pragmatics development might facilitate children to reason about language use on a more sophisticated level of pragmatic interpretation, that, in turn, might be related to, or even based on some form of higher-order mental state understanding (Rakoczy & Oktay-Gür, 2020).

To follow the aim of this dissertation to clarify the developmental trajectories of ToM and ways to validly test these abilities in development, the following research questions need to be addressed in light of the considerations summarized above: First, what are the pragmatic tasks factors that lead to children's failure in the TB task around age four? Second, how does the recovery of performance come about at the end of the U-shaped curve? Do seven to ten-year-old children overcome their intermediate difficulty with TB tasks as the effect of new advances in their pragmatic understanding which might be related to general underlying developmental changes in recursive ToM?

The previous section (2.5.1.6 Competence vs. Performance: Test cases) spelled out empirical possibilities to address the first question. To approach the second question, the following section will start by presenting research on continuing development in ToM in order to provide a basis to address open research questions on more complex forms of ToM and potential relationships to progress in TB performance.

## 2.5. Continuing Development in Theory of Mind

Even though pragmatics accounts on TB performance are compatible with a four-year-revolution to meta-representation ToM, children have certainly not achieved adult-like meta-representational capacities at this age. More complex forms of ToM develop in more protracted ways than basic ToM. In continuing development, children's ToM shows advances in two ways. First, children become able to engage in higher-order mental state reasoning ("A thinks that B thinks that…"), also referred to as *Higher-order ToM*. Second, children learn to apply mental state concepts in flexible ways and become proficient in interpreting complex social situations including, for example, nonliteral speech acts such as irony or jokes (Happé, 1994; Peterson et al., 2012; Peterson & Wellman, 2019; Wellman et al., 2011), recognizing social faux pas (Baron-Cohen et al., 1999), and interpreting complex emotions (Baron-Cohen et al., 2001). This multidimensional group of applied, complex ToM skills, is summarized as *Advanced ToM*[4] (Osterhaus et al., 2022; Wellman, 2018). Tasks to measure Higher-order ToM and Advanced ToM can radically differ from each other. The two following sections describe these two sophisticated forms of ToM, exploring how they can be measured and summarizing their development during middle childhood.

### 2.5.1. Higher-order Theory of Mind

Reasoning about how people represent the world often provides a useful framework to explain and predict their behavior, and how they (*physically*) act toward objects and other people. However, reasoning about representations of the physical world cannot entirely capture the complexity of human *social* interaction (Perner & Wimmer, 1985). Social interaction also often takes into account what other people think about other people's thoughts (second-order beliefs) (Perner & Wimmer, 1985). Over the course of development, such advanced forms of mindreading beyond the understanding of first-order FB ascriptions emerge. Understanding the recursive

[4] *Higher-order Theory of Mind* (or *recursive Theory of Mind;* Osterhaus & Koerber, 2022) is usually considered as one part of *Advanced Theory of Mind* abilities (e.g., Osterhaus et al., 2016). To better distinguish the narrow construct of explicit, higher-order mental state reasoning and the supraordinate construct of applied skills that base on (higher-order) mental state reasoning, they are referred to *Higher-order Theory of Mind* and *Advanced Theory of Mind* respectively (Osterhaus & Koerber, 2022).

nature of mental states enables advanced forms of flexible and higher-order mindreading in which additional levels of mental states can be represented (Perner & Wimmer, 1985). These complex and higher-order forms of reasoning emerge by the recursive embedding of meta-representations. In this way, the level of recursion of every higher-order mental state ascription can be determined (first-order: "A thinks that p", second order: "A thinks that B thinks that p", third-order "A thinks that B thinks that C thinks that p", etc.). Also, higher-order ToM can involve recursive thoughts about each other (in the sense that "A thinks that B thinks that A thinks that p") and can, of course, include people's thoughts about their own thoughts ("She thinks that I think that he thinks that p"). In this way, meta-representations can be recursively embedded within each other virtually ad infinitum.

While children – according to the standard picture of ToM development – solve tasks measuring first-order belief ascription around age four in the course of a major conceptual transition to meta-representational thought, children sophisticatedly attribute second-order mental states ("A thinks that B believes that p") only after a developmental delay. Children's understanding of second-order beliefs is usually tested in an adaption of the change-of-location task (Perner & Wimmer, 1985). In the classic storyline, two protagonists together witness an object in a location (*e.g.*, the ice-cream van in the park). Then, protagonist A leaves the scene. In her absence, protagonist B observes the object being moved to a different location (*e.g.*, the ice-cream van drives from the park to the school). Unbeknownst to protagonist B, protagonist A finds out about the location change. Thus, both protagonists happen to have a TB (*i.e.*, knowledge) about the object's location. However, protagonist B falsely believes that protagonist A still thinks that the object is in location 1. The test question then refers to protagonist B's belief about protagonist A's belief-based action "Where does protagonist B think that protagonist A looks for the object?" (to which the correct answer is location 1 as protagonist B is unaware that protagonist A has seen the object's move) (Perner & Wimmer, 1985). When memory demands are reduced to a minimum, children start to solve these kinds of second-order FB tasks at ages five to six (Miller, 2009; Sullivan et al., 1994).

More interactive implementations of second-order ToM tasks, for example, include peer coordination. In a study by Grueneisen and colleagues (Grueneisen et al., 2015) children had to infer their partner's game decision based on a second-order FB ("She does not know that I know that p. She thinks that I falsely believe that q") and had to adjust their own move accordingly. In accordance with results from classic tasks (Perner & Wimmer, 1985; Sullivan et al., 1994), six-year-olds demonstrated their capacity to use such second-order mental state attributions to successfully coordinate with their peer without any communication (Grueneisen et al., 2015; see also Wyman et al., 2013).

While there is a solid basis of research that shows that children solve second-order ToM tasks with a developmental delay of approximately two years after first-order tasks, very little is known about children's development of higher-order ToM beyond the second-order recursion. Decades of research concentrated on first- (and, partially, second-) order FB understanding in preschoolers. In recent years, ToM research primarily focused on existing new looking-time studies with infants (*see* section 2.4 Underestimation claim: Implicit Theory of Mind) leading to an enormous imbalance in empirical quantity. A striking gap in the literature, thus, concerns developments in ToM during middle childhood and adolescence (Hughes, 2016).

Liddle and Nettle (2006) developed a task in which children heard storylines involving multiple characters. After listening to the storylines, children were asked to judge which of two statements about the characters is correct with regard to the storyline (Liddle & Nettle, 2006). The two statements contained mental state ascriptions from the first to fourth order of recursion. Children aged ten and eleven performed above chance level in first- to third-order ToM tasks but at chance level in a fourth-order ToM task. Osterhaus and colleagues used the same storylines from Liddle & Nettle (2006) and found higher-order mental state reasoning in slightly younger children. In these studies children aged seven and older solved third-order FB test questions correctly (Osterhaus et al., 2016; Osterhaus & Koerber, 2021) while performance in the fourth-order FB task was around the chance level until age ten (67% of the ten-year-olds answered correctly in the fourth-order task, no results of statistical chance-level comparisons provided; Osterhaus et al., 2016).

Most adults, in contrast, are usually able to solve ToM tasks until fifth order of recursion, and, a minority of them, even sixth to eighth-order tasks (Henzi et al., 2007; Kinderman et al., 1998; Oesch & Dunbar, 2017; O'Grady et al., 2015; Osterhaus et al., 2016; Stiller & Dunbar, 2007; but see Wilson et al., 2022). Thus, evidence suggests that higher-order ToM develops gradually over childhood into adulthood in more protracted ways than first-order ToM. Due to a lack of research on such higher-order forms of ToM during middle childhood, the specific developmental trajectories are not yet examined in all detail.

Beyond such explicit, higher-order mental state ascriptions, children also learn to flexibly apply (higher-order) mental state concepts during middle childhood. This facilitates lots of abilities in the scope of social cognition often summarized under the umbrella term *Advanced ToM.*

### 2.5.2. Advanced Forms of Theory of Mind

During middle childhood, children learn to apply mental state concepts in a flexible way and, consequently, become more proficient in interpreting complex social situations in daily life (Apperly, 2012; Devine et al., 2016; Osterhaus & Bosacki, 2022). These age-related developments in applied mindreading abilities that occur after preschool are summarized as *Advanced ToM* (AToM) (Devine & Hughes, 2016). Definitions of AToM are diverse (*see e.g.*, Wellman, 2018), and rather result from the operationalization of tasks that are used to measure developmental and individual differences in ToM abilities in older children and adults (Osterhaus & Bosacki, 2022). These tasks, however, vary considerably with regard to the specific skill assessed, the stimuli presented, and the response format used (Devine & Lecce, 2021). A lack of consensus remains as to which extent the different aspects of AToM cohere (Warnell & Redcay, 2019), and as to which of these tasks validly and reliably assess such a multidimensional concept of social cognitive skills (Devine & Hughes, 2016; Osterhaus & Bosacki, 2022; Weimer et al., 2021).

In this section, I will provide an overview of the constructs and abilities involved in AToM testing. These tasks are used to measure various abilities related to social understanding, such as emotion- and mental-state recognition, non-literal speech acts, and social cognition (Baron-Cohen et al., 1999; Baron-Cohen et al., 2001; Happé, 1994; Osterhaus et al., 2016; Osterhaus & Bosacki, 2022; Osterhaus & Koerber, 2021; Winner & Gardner, 1993).

In particular, I will focus on non-literal speech acts and the tasks used to test these abilities in middle childhood and provide an overview of the results of studies that have examined children's developing performance in these tasks (for comprehensive reviews, *see e.g.*, Devine & Lecce, 2021; Osterhaus & Bosacki, 2022).

One group of tasks tests for participants' social understanding in various situations. For example, they are asked to interpret social situations in silent films (Devine & Hughes, 2013) or scenarios that involve ambiguous social interactions (Bosacki, 2000). In other tasks, participants are required to detect social transgressions, such as faux pas (Banerjee et al., 2011; Baron-Cohen et al., 1999; Happé, 1994). To recognize faux pas as such, participants need to track diverging epistemic states of the involved characters, recognize the impact of a respective statement or action, and integrate their own knowledge about context-sensitive social norms (Baron-Cohen et al., 1999).

A second group of tasks targets children's emotional and mental state recognition. One widely used measure is the *Reading the Mind in the Eyes* Test which requires participants to recognize emotions in photographs of other people's faces, or more precisely, in their eye area (Baron-

Cohen et al., 2001). Participants are asked to match each photograph with a description of mental states (*e.g.*, “annoyed”). This procedure aims to test an affective component of ToM that involves recognition of facial emotional expression and attribution of mental states (Baron-Cohen et al., 2001).

In general, children become more proficient in these tasks of the two groups that cover aspects of social understanding and emotional and mental state recognition in middle childhood (*see* e., Osterhaus et al., 2016). More specifically, however, children’s performance varies considerably within age groups in middle childhood and even early adolescence (*e.g.*, Devine & Hughes, 2016). This substantial variance makes it difficult to make general statements about AToM skills of children in certain age groups (Devine, 2021). That, however, indicates that these tasks can capture individual differences in AToM that are not particularly related to specific age groups (Lecce, 2021). Therefore, these tasks are often used to assess relations between AToM and other variables, such as academic achievement (*e.g.*, Lecce, 2021) or peer relations (Fink, 2021).

A third group of AToM tasks assesses children’s pragmatic language understanding (Dryll, 2009; Filippova & Astington, 2008; Happé, 1993, 1994; Norbury, 2005; Pexman & Glenwright, 2007; Szücs, 2014; Winner et al., 1988; Winner & Gardner, 1993). Pragmatic abilities generally involve comprehending and producing speech acts and discourse beyond its mere literal meaning. To understand most speech acts as they were intended by the speaker (*i.e.,* the speaker’s meaning), additional information besides the literal meanings of the words (sentence meaning) needs to be considered. These include, but are not limited to information about who made an utterance, in which context, and against the background of which conversational rules. Taking into account all this information, the recipient aims to figure out the speaker’s intentions underlying the speech act in question (Searle, 1969). Pragmatics thus involve some form of applied ToM (Happé, 1994; Matthews, 2014).

Ironic and metaphorical utterances are prototypical examples of non-literal speech acts: the speaker does not want the recipient to take her utterance literally. To interpret utterances involving irony and metaphors, the recipient ascribes complex intentions to the speaker (Filippova & Astington, 2008; Happé, 1993). In tasks that test children’s understanding of non-literal speech acts, children are presented with very short storylines including non-literal utterances. Children are then asked to explain these utterances. One typical example is Happé’s *Strange Stories* Task (Happé, 1994) in which children are asked to explain non-literal utterances, such as metaphorical and ironic utterances. Usually, neurotypical children solve tasks that require an understanding of metaphors during school age or even preschool age (Lecce, 2021; Matthews, 2014; Pouscoulous & Tomasello, 2020). In contrast, evidence from tasks with ironic utterances suggests that children develop an understanding of irony only later. In neurotypical development, children come to solve these tasks during school age between six and 13 years of age, depending on the kinds of measures used (Matthews, 2014). The direct relationship between non-literal language understanding and ToM in development has been subject to empirical investigations over the last few years (Del Sette et al., 2020; Lecce et al., 2019; Norbury, 2005). The developmental associations between ToM and the understanding of metaphors are still discussed controversially. In contrast, results on the relation of irony comprehension and ToM in development are more conclusive: children's performance in tasks that require irony comprehension correlates with their second-order FB understanding (Angeleri & Airenti, 2014; Filippova & Astington, 2008).

In addition to metaphorical and ironical utterances, Happé's *Strange Stories* Task (Happé, 1994) incorporates other speech acts that require the ascription of more complex intentions, such as bluffs, double-bluffs, and white lies. In one such storyline containing a double-bluff, a man tells the truth to deceive his interlocutor as his interlocutor expects him to lie. This situation was interpreted by neurotypical children on a third-order theory-of-mind level (*e.g.*, "he knows that they think that he will lie") (Happé, 1994) around the age of seven (Osterhaus & Koerber, 2021).

#### 2.5.2.1. Role of Complex Theory of Mind in the True Belief Task

The previous section presented research on continuing development in ToM to provide a basis to investigate complex forms of ToM, in general, and, more specifically, to examine potential relationships of such complex forms of higher-order ToM to progress in TB performance.

The pragmatics analysis is presented in section 2.4.2.5. "Performance Limitation Account: Pragmatics" predicts that children are pragmatically confused by the peculiarity of the TB test question. A related question that needs to be investigated to approach the underlying mechanisms of TB failure is: How do children come to solve the TB task at the end of the U-shaped curve? Applying the logic of pragmatic tasks demands, children are expected to answer the TB test question correctly once they have made further developmental progress in their pragmatics understanding. Developmental advances in pragmatics might enable children to follow more complex forms of speech acts. In the TB task, this might enable them to interpret the TB test question on a higher, more advanced level of pragmatic interpretation. Consequently, they might become able to resolve their pragmatic confusions about the academic, trivial test question in the TB task (Rakoczy & Oktay-Gür, 2020).

This approach predicts that children should become able to solve the TB task at the end of the U-shaped curve when they become able to reason flexibly on higher-order levels of pragmatic interpretations. Based on this idea, further developmental progress in more flexible forms of pragmatic understanding would determine the end of TB failure around age eight to ten. Developments in higher-order forms of pragmatic understanding might, in turn, be based on the capacity for recursive, higher-order mind-reading. Recursive ToM might enable the ascription of specific higher-order communicative intentions to make sense of the peculiar test question. Children who develop an understanding of more complex forms of pragmatics can either interpret the TB question as an academic test question right away or, if they get confused, ascribe higher-order mental states to overcome their confusion about the task. In one way or the other, children's performance on the TB test question is expected to be related to their general recursive ToM capacities.

After proposing ways to clarify (Ia) the emergence and presenting the current state of research on (Ib) subsequent development in ToM, I will now address questions of potential relations of ToM to (IIa) a general understanding of representations. To this end, I will introduce the idea that Mental Time Travel abilities rely on a similar understanding of representations and more complex forms of Mental Time Travel involve recursive thinking to embed these representations within each other. The purpose of asking the question of how children's developing abilities in ToM relate to other cognitive capacities is to approach the potential underlying foundations of ToM development from a domain-general perspective on ToM.

# 3. Meta-Representation Beyond the Scope of Theory of Mind

Over the last few decades, it has been debated if ToM abilities, the emergence of FB understanding and competencies to apply such mental states, reflect either a domain-specific or a domain-general capacity. In light of strong evidence (summarized in section 2.3.3 General Transition to meta-representational thought) from studies comparing children's emerging understanding of mental and non-mental representations (*e.g.*, Perner & Leekam, 2008), domain-general views have claimed that ToM relies on the more general ability to hold meta-representations (Perner, 1991). Based on this claim, the pressing question arises of how broadly shared this cognitive foundation in other reasoning domains is. Which other abilities rely on the same cognitive foundations of meta-representational thought? One other capacity that recently has been claimed to involve meta-representation is Mental Time Traveling (Redshaw, 2014).

## 3.1. Mental Time Travel

The ability to think and reason about the past (*i.e.*, episodic memory) and the future (*i.e.*, foresight) is often referred to as *Mental Time Travel.* Simple forms of Mental Time Travel seem to be present in multiple species (Cheke & Clayton, 2012; Correia et al., 2007; *but see* Hoerl & McCormack, 2019; Redshaw, 2014). The decisive factor that distinguishes complex Mental Time Travel from such simpler forms is that the individual does not only need to mentally revisit or pre-visit events from the past and the future but also needs to understand how these events relate to the present (Buckner & Carroll, 2007; Schacter et al., 2007; Suddendorf & Corballis, 2007; Tulving, 1985).

### 3.1.1. How to Measure Mental Time Travel

To test for this sophisticated capacity, recent research focused on future-oriented Mental Time Travel, including an understanding of how future events relate to the present. To test this capacity to imagine, entertain, and reason about future events, participants are usually faced with a situation that opens up multiple, still undetermined future possibilities[5]. Importantly, the uncertain future possibilities need to be mutually exclusive future possibilities – only one of two or more possibilities will actually occur in the immediate future. This requires the participant to reflect on how the possible future events related to the present: all possible future events are possibilities of the same present state, i.e., the same referent. Consequently, as soon as one of these possibilities actually occurs, all other possibilities are precluded from being true/real. This feature requires to reflecting on the relation between the actual state and the possible future state (Redshaw, 2014; Redshaw & Suddendorf, 2020).

In the literature, two main groups of tasks were used to test this form of future-oriented Mental Time Travel. The two groups of tasks differ with regard to their structure: Participants either need to prepare for multiple uncertain mutually exclusive future events in parallel or they need to choose a certain option over mutually inclusive but uncertain options. Importantly, these

[5] For an overview of other attempts and their limitations, *see* Gautam et al., 2019; Hoerl & McCormack, 2019

tasks require reasoning about the uncertain future (also referred to as physical uncertainty, Robinson et al., 2006). The outcome of the event which has yet to happen is not yet determined so participants are required to represent possibilities about what could happen in the future[6].

Robinson and colleagues (2006) designed a task in which participants are required to prepare for multiple uncertain mutually exclusive future events in parallel. In their task (Robinson et al., 2006, Exp.1, "Unknowable condition") participants were told that the experimenter was going to draw one block out of a bag that contains both orange and green blocks. The block was then pushed through a door of the respective color (green door or orange door). Participants were asked to prepare trays so that they catch the block that was going to be drawn and pushed through the door. As the participants were asked to prepare before the block was actually drawn from the bag ("Unknowable condition"), the future (*i.e.*, the block being pushed through the green door or the block being pushed through the orange door) was still uncertain and was only going to be determined in future. Children who were able to represent both future possibilities (the block being pushed through the green door or the block being pushed through the orange door) were expected to cover both the green and the orange door with trays (Robinson et al., 2006).

In a structurally similar study by Beck and colleagues (Beck et al., 2006, *Exp. 1 and 2*; "undetermined" trials), participants were required to put out mats to catch a mouse that will come out of either of two sides of a forked-shaped (upside-down y-shaped) slide. When participants were asked to prepare to catch the mouse, the two branches of the forked-shaped slide were blocked by gates. They were told that only one of the two gates will open so that the mouse can pass and go down the respective branch of the slide. Which gate will be open was going to be determined by a card that will be drawn by the child (upside-down). Again, covering all possible exits with mats was interpreted as evidence for future possibility reasoning (Beck et al., 2006).

In a simplified version (Redshaw & Suddendorf, 2016), participants were asked to catch an object with their hands. They were told that the object will be dropped into a forked tube. Only when using both hands simultaneously, participants can cover both lower exits of the tube and make sure that they will definitely catch the object that will come out of either of the two

[6] Crucially, all these tasks need to be distinguished from tasks that require reasoning about possibilities under epistemic uncertainty. These tasks often follow the same logic with the important difference that the outcome of the relevant event has already been determined in the past, yet remains unknown by the observer. In these tasks, successful participants represent possibilities about what might already be the case in the present (*e.g.*, Robinson et al., 2006, *Exp. 1* "unknown condition"). This structural change results in a more complex representational structure: participants need to represent that, in the past, there have been multiple alternatives about the future and that, by now, reality has been determined but remains unknown. Consequently, participants have to represent their own ignorance about reality (*see*, *e.g.*, Robinson et al., 2006; Redshaw & Suddendorf, 2020).

branches. Interestingly, the future event here (*i.e.*, the object coming out on the left or the right exit of the forked tube) is determined exclusively physically. In contrast, in the tasks by Robinson and Beck, the future will be determined by the color of the block that is drawn by the experimenter (Robinson et al., 2006) or the card that is drawn by the child (Beck et al., 2006). These task designs could lead to the impression that the future event is not exclusively determined physically but also intentionally (by the experimenter drawing the block; by the child drawing the card) or conventionally (by the color matching rule of the block and the door or the card and the gate).

In tasks that follow the second task structure participants are required to choose a certain option over mutually inclusive but uncertain options to pass the task and to demonstrate their Mental Time Travel capacities. In another version of the forked slide task (adapted from Beck et al., 2006), participants are told that two objects will be dropped into two slides simultaneously. One of the slides is non-branching (having one exit at the bottom), and the other slide (the forked slide) is branching into two exits at the bottom. Participants are asked to cover only one of the three exits of the slides with a wagon in order to catch an object (Leahy, 2023). Again, the future event for the forked slide (*i.e*., the object coming out on the left or the right exit of the forked slide) will be determined physically. Unlike the other tasks (Beck et al., 2006; Redshaw & Suddendorf, 2016; Robinson et al., 2006), participants in this task were not required to prepare for two options of an uncertain event in parallel (*i.e*., putting out two mats under two exits) but identify the uncertain, undetermined future possibilities (object coming out of either of the two exits of the forked slide) and the certain, determined future event (object coming out of the bottom exit of the non-branching slide). Based on the comparison of the (un)certainty of the future events, participants were required to choose a certain option and cover the exit of the non-branching slide over the uncertain options (one of the exits of the branching slide).

Armed with signatures of sophisticated Mental Time Travel and standard ways of measurement in possibility reasoning tasks, I will present empirical evidence to give a broad overview about the ontogenetic emergence of Mental Time Travel next section.

### 3.1.2. Ontogeny of Mental Time Travel

Dependent on the exact task structure and design, children reliably pass tasks in which they are required to reason about uncertain future events, around the age of four to five years. The majority of results come from tasks in which participants need to prepare for multiple uncertain mutually exclusive future events in parallel. In Robinson and colleagues' block-door-task, participants between four and a half and six and a half years of age succeeded to prepare for this uncertain future event by putting out two trays (under both the green and orange door) in this condition (Robinson et al., 2006). Beck and colleagues' slides task showed similar results. Five- to six-year-old children succeeded to prepare for this uncertain future event by putting two mats under the two exits of the forked-shaped slide. Three- to four-year-olds, in contrast, failed to succeed in this condition. They did not consistently put two mats under the two exits (Beck et al., 2006). In the simplified version by Redshaw and Suddendorf (2016), the majority of four-year-old participants succeeded in reliably covering both exits with their hands from the first trial while younger children and non-human primates were not able to do so. In the task in which participants were required to choose a certain over mutually inclusive but uncertain options, children passed around age five. While four-year-old children failed to consistently put the wagon under the exit of the non-branching slide, children aged five and older succeeded (Leahy, 2023).

In summary, Mental Time Travel emerges from age four to five in its basic form when children start to imagine future events and relate those to the current present. They start to reason sophistically about uncertain future events while children younger than four often fail to consider multiple uncertain possibilities. Instead of preparing for multiple possible events, young children often prepare for only one of the possibilities (*e.g.*, see Leahy & Carey, 2020 for a "one possibility treated as a real" account). Yet, there are, of course, more complex forms of Mental Time Travel and, thus, still much development to come after the age of four to five.

## 3.2. Recursive Mental Time Travel

In more complex and higher-order forms of Mental Time Travel additional levels of temporal representations can be represented by recursive embedding. This allows to remember past moments in which one thought about earlier times or to imagine that in the future one will make plans for the further future. Also, complex Mental Time Travel can involve switching back and forth between past and future perspectives: remembering moments in the past in which one thought about the future or imagining that in the future one will look back on the relative past (Redshaw & Suddendorf, 2020).

One example of such recursive forms of Mental Time Travel is counterfactual reasoning about past events (like in "What if I had not studied psychology?" or "What if I had won the lotto jackpot yesterday?"). One way to conceptualize counterfactual thought is that one has to mentally travel back in time and change aspects of the past in order to imagine how this alternative version of the past (level 1) would have affected the future and thus led to an alternative present (level 2). Based on this conceptualization, it can be argued that counterfactual thought requires a second-order recursive operation as the representation of the alternative present has to be embedded into the representation of the alternative past. Crucially, in this process, the relation between simultaneous representations of the world (actual reality and counterfactual alternative) needs to be represented (Beck & Crilly, 2009; Beck & Guthrie, 2011; Byrne, 2016; McCormack et al., 2018). Actual reality and the counterfactual alternative need to be represented as alternative versions of the very same moment in time that originates from a common past (Beck et al., 2006; Redshaw & Suddendorf, 2020).

### 3.2.1. Counterfactual Reasoning: Measures and Ontogeny

To validly investigate counterfactual reasoning, tasks need to fulfill some stringent desiderata. This is important as tasks need to distinguish true counterfactual reasoning from other, simpler forms like conditional reasoning strategies that mimic counterfactual reasoning abilities under some conditions (*see* e.g., Leahy et al., 2014). First, tasks need to address real-world counterfactual reasoning in contrast to other types of non-actual scenarios, including pretense, fiction, and hypothetical thought. These have been suggested to be different abilities that vary in their representational structure (Beck et al., 2006; Beck & Guthrie, 2011; Rafetseder et al., 2010, 2013).

Second, the joint past of actual reality and the counterfactual alternative constraints the construction of the counterfactual as *Nearest Possible World* (*see* Rafetseder et al., 2013). The counterfactual past needs to equal the actual past except from the relevant counterfactual change. Consider the following situation: you just cleaned the kitchen floor, so that it shines brightly when your daughter, Carol, comes home from the playground still wearing her muddy shoes and runs over to you so that the kitchen floor is littered with her footprints (Harris et al., 1996). Now imagine, what if Carol had taken off her shoes before she went into the kitchen? Would the kitchen floor be clean or dirty? In the construction of the counterfactual past in which Carol had put off her shoes, everything needs to equal the actual past (*i.e.*, you just cleaned the kitchen floor, Carol comes home from the playground wearing her muddy shoes) only differing in terms of the change mentioned in the question (Carol puts her shoes off) (Harris et al., 1996).

To systematically test for real-world counterfactual reasoning about the nearest possible world that can be distinguished from simpler forms of reasoning, Rafetseder and colleagues introduced tasks that require counterfactual reasoning about so-called *overdetermined events* (Rafetseder et al., 2013). In their adaption of the “dirty shoe story” (by Harris et al., 1996), Carol brings her friend Max. Now, two children rather than only one child walk over the clean kitchen floor with their muddy shoes on so that the kitchen floor becomes dirty (Rafetseder et al., 2013). After seeing the storyline, participants were asked whether the floor would have been dirty if one of the children say, Carol, had taken her shoes off. In this scenario, the outcome (the dirty kitchen floor) is overdetermined (the floor gets dirty by two causes: Max and Carol). Thus, the possibility that participants solve the counterfactual test question with simpler conditional reasoning (*e.g.*, hypothetical thinking or if-then-principles: if shoes are taken off, the floor is clean) is reduced. The results of studies using such over-determination scenarios showed surprisingly low levels of performance. Five- to six-year-olds answered correctly to 18% of the test questions, children as old as nine were correct only 53% of the time, while adults and older adolescents performed at or close to the ceiling (Rafetseder et al., 2013).

However, these kinds of storylines and test questions all come with problems (*see* McCormack et al., 2018). One line of concern addresses the high levels of memory and attentional resources that are required to follow the storylines. A second concern is that the scenarios are mentally rather than physically caused (*e.g.*, Carol taking her shoes off) and might, therefore, impact whether other events occur. For example, the counterfactual change (*i.e.*, Carol taking her shoes off) might influence others’ behavior (*e.g.*, Max putting his shoes off, too) (Kratzer, 1981). Lastly, the answer format of the question is usually binary even though the outcome is not. To spell that argument out, let us consider the dirty shoe story once again. In the beginning, the floor is very clean and after Max and Carol walked into the kitchen, it is very dirty. In the study by Rafetseder and colleagues (Rafetseder et al., 2013, *Exp*. 2) this is acted out with two little figurines walking over a plate with dirt on their shoes. In the counterfactual scenario in which Carol takes her shoes off, only Max would make footprints on the floor. In comparison to the very clean floor from the beginning of the story, the floor would be of course still dirty, even if only one person with dirty shoes had walked into the kitchen. In comparison to the actual outcome of the story (very dirty floor because of footprints of two persons), however, the floor would be still dirty but cleaner if only one person with dirty shoes had walked into the kitchen. Yet, the answer format does not allow for such a gradual emphasis on the state of affairs.

In summary, the combination of these problems might lead to false negatives in counterfactual reasoning tasks in young children (*see* e.g., McCormack et al., 2018). At the same time, the task format might also allow for false positives. The use of familiar situations in these storylines might increase the probability that participants answer the counterfactual reasoning test questions based on simpler conditional reasoning (*e.g.*, generalized if-then-principles; McCormack et al., 2018). All in all, it is unclear whether these tasks validly test the competence of true counterfactual thinking. On the basis of this critique, McCormack and colleagues designed a task that avoids all of these problems. In their pig-disc task (here: “doubly determined” condition), children saw an apparatus with two ramps on which discs are rolling down and which both end at the center of the apparatus. One ramp – the fast one – leads directly to the center, the other slow ramp has a zig-zag shape. As soon one disc reaches the center of the apparatus, it hits a toy pig which consequently falls down from the apparatus. In the relevant test trials, participants saw two discs rolling down the ramps simultaneously with the effect that the disc on the direct ramp hits the pig first (pre-empting the effect of the other, slower disc). After watching this scenario, participants were asked to answer the counterfactual test question about this pre-empted (overdetermined) event. McCormack and colleagues systematically compared additive and subtractive counterfactual reasoning questions. While subtractive scenarios require one to mentally undo something that has in fact occurred, additive counterfactual scenarios require one to mentally add something to what actually happened (“If I had not rolled the [fast disc] down that time, would the pig have fallen?” for the subtractive trial and “If I had put the peg in here (*i.e*., a peg that blocks the runway of the direct ramp and therefore prevents the first disc from rolling down) that time would the pig have fallen?” for the additive scenario).

The advantages of this task are that it introduces a new, unfamiliar scenario about a physical causation event with a clear binary outcome (either pig has fallen or has not fallen) that comes with rather low memory demands. In contrast to earlier studies testing counterfactual reasoning about over-determined events, the results show that children starting from six years were able to answer counterfactual test questions about doubly-determined scenarios correctly. Four- to five-year-olds, however, answered incorrectly (McCormack et al., 2018).

In summary, success in tasks testing counterfactual thinking about past events as a form of recursive Mental Time Travel comes with a developmental delay after reasoning about future possibilities. Dependent on the exact task structure and design, children reliably answer counterfactual test questions correctly at the ages of six to nine.

### 3.2.2. Counterfactual Emotions: Measures and Ontogeny

Around the time children engage in counterfactual reasoning, children typically also begin to experience counterfactual emotions such as regret and relief (McCormack et al., 2018, 2019; O'Connor et al., 2012; Uprichard & McCormack, 2019; Van Duijvenvoorde et al., 2014; Weisberg & Beck, 2010). These emotions arise from a comparison that builds on second-order Mental Time Travel. Experiencing regret or relief requires the agent to realize that in the past (level 1) different options for the relative future (level 2) were available (Hoerl & McCormack, 2019). By comparing the true present and an alternative present, the actual state of the world is evaluated as better (relief) or worse (regret) than the imagined counterfactual.

To measure the ability to experience regret or relief in development, most of the studies followed a similar paradigm. Children are asked to choose one of two boxes in order to win its content. Children then learn about the prize they won but also about the content of the unchosen box which contains a better (regret trials) or worse (relief trials) prize. Regret and relief are assessed by asking children to report on their emotions once they learned about the prize in the chosen box and again once they learned about the prize in the other box. Children are expected to report feeling less happy (regret trials) or even happier (relief trials) than they felt before when they only knew about the actual prize. Children come to report feeling regret around age six, but the exact age of onset varies between studies (age six to seven years in O'Connor et al., 2012; 2015; Burns et al., 2012; age seven to eight years in Van Duijvenvoorde et al, 2013; age four to five years in Weisberg and Beck, 2012; age nine years in Rafetseder & Perner, 2012).

Adding yet another layer of recursion, the anticipation of counterfactual emotions constitutes a case of third-order Mental Time Traveling. Anticipating regret or relief requires the agent to imagine that they will learn in the future (level 1) that – in the relative past (level 2) – better options for the relative future (level 3) were available.

To test the ability to anticipate regret, the two-boxes task was adapted in different ways. For example, children were asked what they would like to find in the unchosen box (Guttentag & Ferrell, 2008) or what might be in the other box that would make them feel sadder (McCormack & Feeney, 2015, *Exp. 1*). In other tasks, children were first asked how they felt about choosing their box and after that, how they would feel about it if they would find out that there was a larger prize in the other box that they have not chosen (McCormack & Feeney, 2015, *Exp. 2&3*). Similarly, to test the ability to anticipate relief children were first asked how they felt about choosing their box and after that, how they would feel about it if they would find out that there was a smaller prize in the other box that they have not chosen (McCormack & Feeney, 2015, *Exp. 2&3*). The ability to anticipate to experience counterfactual emotions such as relief or regret in the future seems to emerge even later in development. Only from around age eight or later, children are able to report that they will have negative emotions if they will learn that the option, they chose turns out to have been worse compared to other options (Guttentag & Ferrell, 2008; McCormack & Feeney, 2015; O'Connor et al., 2015; Rafetseder & Perner, 2012).

# 4. Relation of Theory of Mind and Mental Time Travel

The summary of evidence on the development of both ToM and Mental Time Travel presents similar stepwise developmental progress in the two abilities. Empirical evidence suggests that children first become involved in first-order Mental Time Travel at around age four to five, when they are able to reason about future possibilities (Beck et al., 2006; Leahy, 2023). With a developmental delay of one or two years, children start to think counterfactually about past events as an instantiation of second-order Mental Time Travel around six years of age (McCormack et al., 2018). However, it is not until a further developmental delay that children begin to engage in higher-order Mental Time Travel beyond the second-order of recursion, which typically occurs in middle childhood. For instance, the ability to anticipate experiencing counterfactual emotions such as regret in the future does not typically emerge until around eight years of age (McCormack & Feeney, 2015).

The standard picture[7] of ToM development shows a similar stepwise developmental progress is observed in children's ToM abilities. From around age four, children come to solve FB and conceptually related tasks that require first-order meta-representational ToM (Wellman et al., 2001). With a developmental delay of one to two years, children start to attribute second-order mental states at the age of five to six and higher-order mental states above second-order even later, in middle childhood (Liddle & Nettle, 2006; Sullivan et al., 1994). The developmental trajectories of first- to higher-order Mental Time Travel and ToM seem to follow a parallel pattern. This raises the question of whether these parallel developmental trajectories are a mere coincidence or whether they are actually based on some form of relation or shared underlying cognitive foundations (Gautam et al., 2019).

In line with domain-general views on ToM, one possibility to explain this developmental parallel is that both ToM and Mental Time build on a common capacity for simulation and projection as shared cognitive foundations. Explicit ToM requires simulation to shift the self-perspec-

[7] As discussed in Section 2.4. Challenging the Standard Picture, the claim of an emergence of ToM has been called into question and thus needs to be re-addressed in empirical work.

tive to an alternative perspective. Similarly, hypothetical thinking about the future requires projection to shift from the current reality to the alternative, imagined future point in time (Buckner & Carroll, 2007). Another (not necessarily mutually exclusive) possibility is that both ToM and Mental Time Travel not only share overlapping processes of projection and simulation in general but build specifically on the same underlying cognitive ability to hold meta-representations (Gautam et al., 2019; Redshaw, 2014; Tulving, 2005). In the case of (explicit) ToM, meta-representation allows representing someone's mental state as an alternative representation (Perner, 1991; Pylyshyn, 1978; Suddendorf, 1999). Different claims have been made to what extent Mental Time Travel relies on the same meta-representational capacity. One proposal is based on the idea that Mental Time Travel involves a form of (autonoetic) representation of one's own past or potential future experiences and perspectives and consequently on a form of meta-representation (Tulving, 2005; *see also* Perner & Ruffman, 1995; Perner et al., 2007). Another proposal claims that meta-representation is needed to represent how various temporal perspectives relate to what they represent: specific representational relations of the entertained content to present or future states of affairs in the world have to be processed (*see* Figure 1). Some minimal meta-representational grasp is required to understand the way temporal representations relate to the world (Bieri, 1986, *see also* Redshaw, 2014).

So far, indirect comparisons of ToM and Mental Time Travel draw an ontogenetic picture of a similar age of emergence for the two capacities (Beck et al., 2006; Leahy, 2023; Redshaw & Suddendorf, 2016; Robinson et al., 2006; Wellman et al., 2001). A recent study tested children's ability to perform both future Mental Time Traveling (in a modified version of the slides task) and a FB task. The results showed no difference in the age of success between the two tasks (Halberda et al., 2023, *Exp. 4*; *but see* Immel et al., 2022). In general, this developmental parallel is compatible with domain-general ideas, but it does not provide conclusive evidence for shared foundations and related development. Therefore, more systematic and direct evidence from comprehensive correlational studies is needed to show associated development, despite both abilities being mastered around the same age.

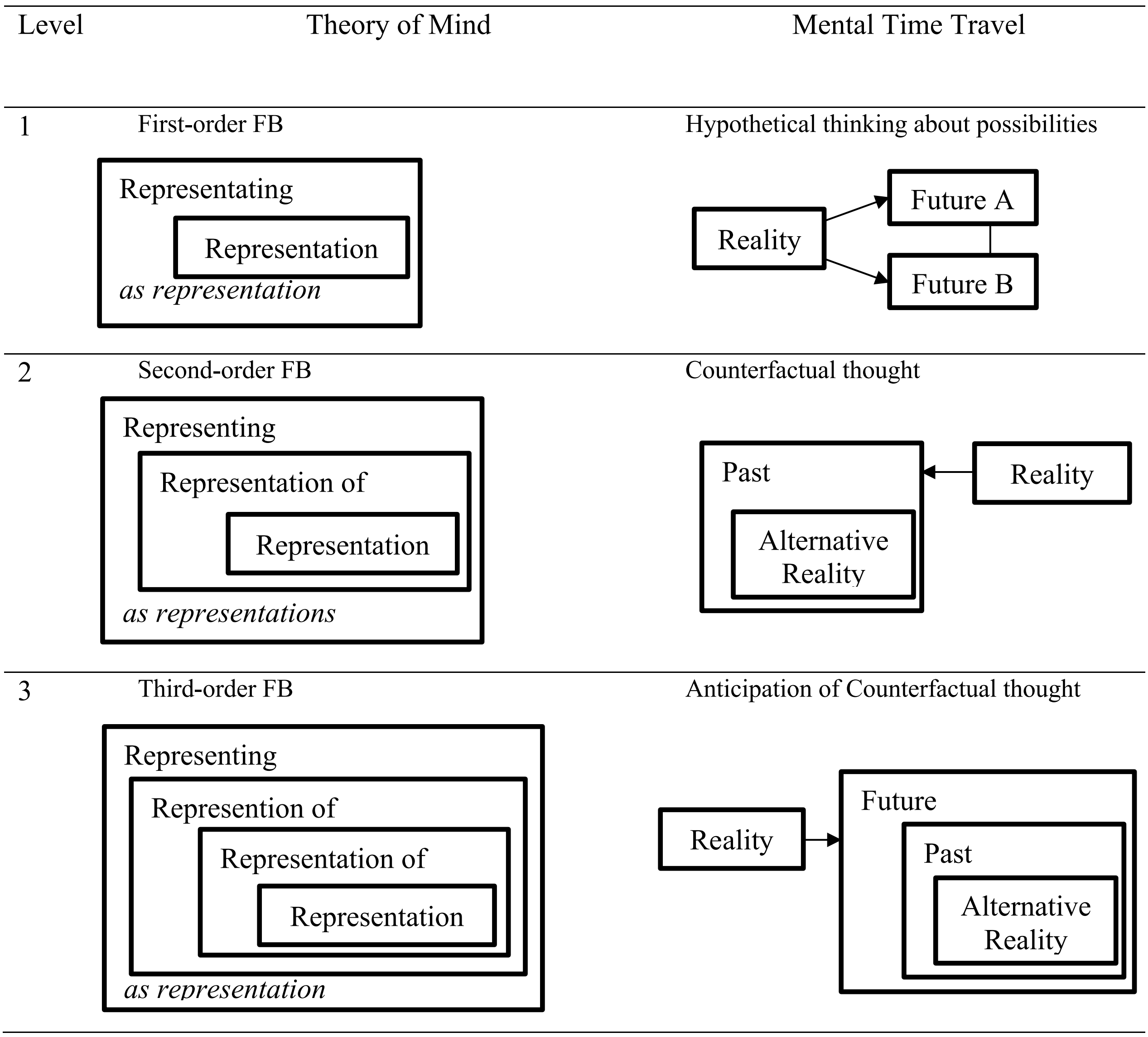



**Figure 1.** Levels of representation in first-, second and third-order Theory of Mind and Mental Time Travel

## 5. Conclusion and Open Questions

Fully-fledged ToM describes the ability to attribute mental states to oneself and to others and to predict and interpret people's behavior based on their mental states (Premack & Woodruff, 1978). At the conceptual core of ToM lies meta-representation (Perner, 1991), the capacity to represent others' (or one's own former) subjective representations of the world even when differing from the own current representation of reality. The litmus test for meta-representational ToM is the FB task in which children need to grasp misrepresentations aiming to demonstrate their understanding of the subjective notion of mental states (Wimmer & Perner, 1983). Children come to pass the FB task around age four in the course of a major conceptual transition to meta-representational thought (Perner, 1991; Wellman et al., 2001). Around this time, they also begin to solve other conceptually related tasks that differ with regard to their structure but all require meta-representational thinking (Perner & Roessler, 2012).

This view of the onset of meta-representational ToM around age four has recently been questioned by a set of surprising findings: Once children master FB tasks, they begin to fail TB control tasks and only in the age of eight to ten, they succeed again (Oktay-Gür & Rakoczy, 2017). Competence limitation accounts claim that the findings of TB failure show that the FB task overestimates children's ToM competencies. Children may use simple heuristics rather than true meta-representation at the age of four. Pragmatic performance limitation accounts, in contrast, claim that these findings do not show any conceptual limitations in children's (first-order) ToM, but rather reflect children's confusion in view of pragmatic task demands in TB tests. To empirically test competence versus pragmatic performance accounts against each other, clear test cases are needed in which pragmatic tasks factors of TB tests are systematically modified. While evidence for actual competence limitations would cast doubt on the four-year revolution, evidence for pragmatic performance problems in the TB task would be compatible with the widely held view of the conceptual transition to (first-order) meta-representational thought around age four.

In subsequent ToM development, more complex and higher-order forms of mental state reasoning emerge by recursive embedding of representations. Children become able to attribute second-order mental states (such as beliefs about beliefs) around age six and use their mental state concepts to interpret social situations and pragmatic speech acts. However, little is known about the exact developmental trajectories of mental state reasoning beyond the first and second-order of recursion and potential relations to pragmatics in the TB task.

According to domain-general views on ToM, emergence and development of ToM relies on an understanding of representations as representation and therefore is marked by the cognitive capacity to recursively embed representations within each other. This capacity is not limited to ToM alone, but is shared with other abilities (Perner, 1991; Perner & Roessler, 2012). Mental Time Travel is claimed to require similar representational capacities to represent that and how past or future events relate to the present state. Recursive iteration can generate more complex and higher-order forms of Mental Time Travel in which additional levels of temporal representations can be represented (Gautam et al., 2019).

Based on these considerations, the following sets of open research questions arise:

(Ia) When does meta-representational ToM emerge? Does children's TB failure around age four provide evidence for limitations of children's ToM capacity or rather performance limitations?

(Ib) How does recursive meta-representational ToM develop throughout childhood and how can this development be measured?

Building on that, it can be addressed how the emergence and subsequent development in ToM relate other to cognitive abilities:

(IIa) Is ToM emergence and higher-order ToM development related to developmental progress in a general (recursive) understanding of representations?

(IIb) Is higher-order ToM development related to the ability of recursive thinking in development?

## 6. Aim of Dissertation

In light of these open questions, I aimed to examine how meta-representational ToM emerges, how recursive forms of ToM evolve throughout childhood, and how its development is related to other cognitive and representational abilities. The overarching aim is to contribute to the understanding of the ontogenetic development of ToM with regard to the debate on domain-general views.

This aim was addressed in three empirical projects. Due to the restrictions of the Covid-19 pandemic, data collection for most of the projects of the dissertation was conducted remotely. Thus, an additional preliminary project was conducted to ensure that online testing is a valid alternative to classical in-person test settings. In this preliminary project, three to four-year-olds' ToM performance in the new online implementation of the false and TB task was compared to in-lab settings.

To approach research questions (Ia) When does meta-representation ToM emerge, and (IIa) Is ToM emergence related to developmental progress in a general understanding of representations, the first main Project compared children's performance in tasks that systematically varied in their representational and pragmatic structure. Children's performance in the false and TB task was compared to parallel tasks that required representational understanding of non-mental, external representations. The following pattern of results was expected for this comparison. First, based on the widely held view of a joint onset of meta-representational thought, parallel and correlated development in mental and non-mental representational tasks was expected. Second, testing the pragmatic performance limitation account for TB failure, dissociated development of tasks deviating in their representational or pragmatic structure was expected: children were expected to show problems in the TB task but not in a task with reduced pragmatic demands.

To address research questions (Ib) How does recursive meta-representational ToM develop throughout childhood and how can this development be measured, and (IIb) Is higher-order ToM development related to the ability of recursive thinking in development, the second project investigated subsequent development of more complex forms of recursive meta-representational thinking in ToM. In addition to their recursive ToM abilities, children's general abilities for recursive thinking were measured. It has been argued that the conceptual core of ToM is meta-representational thought which, in turn, requires a recursive operation: holding a representation of a representation, *i.e.*, meta-representation. More complex, higher-order forms of ToM build on that by recursively embedding meta-representations within each other. In three studies, six- to ten-year-old children were tested in both recursive ToM from first- to fifth-order of embedding and in their general recursive thinking abilities operationalized as syntactic recursive abilities. Based on the idea that meta-representational ToM requires recursive operations, performance in these two tasks was expected to be associated with development. This comparison, hence, ought to shed light on potential underlying cognitive foundations of recursive ToM development in recursive thinking.

To expand on the research question (IIa) which explores the relationship between not only the emergence of ToM but also the development of higher-order ToM and a general understanding of representations, the third project compared the developmental progress in recursive ToM with children's recursive Mental Time Travel abilities. It has been argued that sophisticated Mental Time Travel requires a representation of how past or future events relate to the present state, and thus is based on some form of projection or even meta-representation. From this domain-general view, children are expected to show parallel and associated development in recursive meta-representational thought in the ToM and Mental Time Travel domain. The comparisons of developmental progress in recursive, higher-order forms of ToM and children's recursive Mental Time Travel abilities addressed the question of an underlying capacity for recursive representational thought as an underlying factor of ToM emergence and development.

# 7. Summary of Empirical Findings

In the following section, I will summarize the three main projects that I conducted in the course of my dissertation: (1) Schidelko, Huemer, Schröder, Lueb, Perner, & Rakoczy (2022) – Why do children who solve FB tasks begin to find TB tasks difficult? A test of pragmatic performance factors in ToM tasks, (2) Schidelko, Proft, & Rakoczy (2022) – How do children overcome their pragmatic performance problems in the TB task? The role of advanced pragmatics and higher-order ToM, and (3) Schidelko, Baumann, Proft, & Rakoczy (2023) – Do Theory of Mind and Mental Time Travel abilities build on joint cognitive foundations? A Comparison of development of recursive Theory of Mind and Mental Time Travel. Additionally, I will briefly summarize the preliminary project that was conducted in order to validate the methods that were used in the main Projects: Schidelko, Schünemann, Rakoczy, & Proft (2021) – Online testing yields the same results as lab testing: A validation study with the FB task. I will describe the experimental design and the main results. For further details regarding participants, procedure, analysis, and results, please refer to the original manuscripts (Appendices A-D).

## 7.1. Pre-Project: Schidelko, Schünemann, Rakoczy, & Proft (2021)

Due to the restrictions of the COVID-19 pandemic, routines of classic methods of data collection in interpersonal, face-to-face settings between an experimenter and a child were disrupted. Data for most of the projects of my dissertation needed to be collected in a new, contact-free way. Thus, I established an online paradigm for data collection with children. In this new set-up, the experimenter meets the child in a video conference and interacts with them via video and audio streaming. This shift from classic in-person test settings to remote test settings changed the parameters of interaction between the experimenter and the child enormously, especially concerning the pragmatic context. As previous studies showed the important influence of pragmatic factors on children's performance in ToM Tasks, the preliminary project aimed to validate the new method of data collection in the first step. To evaluate the data quality and comparability of results, the study investigated whether the results found in the in-lab face-to-face settings conceptually replicate in analogous studies implemented in the new parallel online paradigm. For the comparison of the two test settings, the standard change-of-location FB task (Study 1 and 2) and the TB version of the task (Study 2) were administered. Children's performance on these well-established tasks in the new online test setting was compared to the data of matching tasks collected in the context of in-lab studies. In the in-lab setting, tasks were acted out by the experimenter with little figurines. In the online implementation, the experimenter presented the same storyline with the help of animated videos and slides. Results revealed that the typical developmental pattern of performance in these tasks found in in-lab studies could be replicated with the novel online test procedure and the test settings did not impact the performance. Three to four-year-olds' performance in the online version of the FB task increased while performance in the TB task decreased with age. These results suggest that the new method of online data collection with children provides a valid alternative to classical in-person test settings. Data collection for the main projects followed this setup.

## 7.2. Project 1: Schidelko, Huemer, Schröder, Lueb, Perner, & Rakoczy (2022)

The first main project aimed to investigate the emergence of meta-representational thought. Children's failure in the TB control condition cast doubt on the onset of meta-representational ToM around age four. Previous studies, however, provided preliminary evidence for performance rather than competence limitation in the TB task. Children might fail this task because it is pragmatically demanding as it involves an (1) trivial, (2) academic test question (3) that pertains to a subjective mental representation. Clear and stringent evidence for the importance of factor (3) was still missing. Conclusive evidence would strengthen the explanatory power of pragmatic analyses of the TB error and support the claim of the emergence of meta-representational ToM around age four. Therefore, project 1 aimed to provide evidence for both the importance of the third factor on TB performance and for the onset of meta-representational thinking around the age of four. To this end, a clear test case was needed that holds the first two factors constant and manipulates the third factor. This requires a task that involves a (1) trivial, (2) academic test question that, however, (3) does not pertain to a mental but non-mental representation. Based on this idea, children ($N$= 88, 3-6 years) received four tasks: the classic FB and TB task and a new False and True Sign task (adapted from Parkin, 1994). In the Sign tasks, a signpost in a story scenario indicates an object's location either correctly (True Sign condition) or incorrectly (False Sign condition). In close analogy to the change-of-location belief task, the child observes a vehicle (*e.g.*, a tractor) moving from one location (the forest) to a new location (the farm). The signpost between the two locations indicates where the vehicle currently is. In the True Sign version, the signpost veridically indicates that the tractor is first in the forest and is then changed accordingly indicating that the tractor is at the farm. In contrast, in the false sign version, the signpost is not changed and thus becomes a false sign. It indicates that the tractor is in the forest even after it moved to the farm. In both versions of the task, children were asked the test question, “What does the sign say where the [vehicle] is?” The True Sign task, in strong parallel to the TB task, meets factors one and two but not factor 3. It involves an academic test question in a trivial scenario but about a non-mental representation. The False Sign task, in strong parallel to the FB task, tests children’s understanding of the false representation of the signpost – a non-mental representation. Based on claims of a joint onset of mental and non-mental representational thought, children were expected to show correlated performance in the two tasks requiring meta-representational thought: the FB task and the False Sign task. In contrast, they were expected to show performance limitations in the TB task leading to negatively correlated performance with the FB task. Furthermore, they were expected to show no difficulties in the True Sign task.

In line with these predictions, the performance of the two tasks that require mental and non-mental meta-representation correlated positively. Children either failed both tasks or passed both tasks. In contrast, performance in the two conditions of the belief task was negatively correlated. Younger children who failed the FB task passed the TB task. On the other hand, older children who passed the FB task failed to answer the TB test question correctly. Finally, the pattern of performance in the True Sign task deviated from all other tasks. Children across all age groups performed competently in this task so that there was a marked dissociation between TB and True Sign tasks. Only the combination of all three factors (a trivial academic test question regarding an agent’s mental representation), as in the TB task, makes the question pragmatically odd and difficult. A task that poses a trivial academic test question about a non-mental representation as in the True Sign task, however, does not pose any such challenges.

This provides new and clear evidence that the third factor, the TB test question’s reference to an agent’s subjective perspective, is a crucial part of what makes the question pragmatically odd, and thus, difficult for children.

With regard to research questions (Ia) When does meta-representation ToM emerge, and (IIa) Is ToM emergence related to developmental progress in a general understanding of representations, project 1 provides evidence of a general onset of meta-representational thought in ontogeny. Children become able to hold representations about both mental and non-mental representation in parallel during pre-school years. Before this transition to meta-representational thought, children can solve parallel tasks that do not require meta-representational thought. Under certain circumstances, such as when a trivial test question refers to a mental representation, tasks can become confusing for children who transition to meta-representational thought and might require further development to be (re)solved.

### 7.3. Project 2: Schidelko, Proft, & Rakoczy (2022)

Building on the results of a unified emergence of meta-representational thought, the second project investigates the subsequent development of more complex forms of recursive ToM after the four-year revolution. In subsequent ToM development, more complex and higher-order forms of mental state reasoning emerge by recursive embedding of representations. In my second project, I aimed to, first, examine how such recursive forms of meta-representational ToM evolve throughout childhood and second, how its development is related to other abilities that potentially build on (1) recursive forms of ToM or (2) general recursive embedding abilities. Based on results from project 1 that TB performance reflects pragmatic confusion about the test question, it was hypothesized that the recovery of TB performance around age eight to ten reflects underlying developments in children's growing pragmatic awareness. Pragmatics understanding involves the ascription of complex intentions and beliefs and is therefore suggested to be an instantiation of advanced ToM. Therefore, the set of studies tested whether developmental progress in recursive ToM relates to children's recovering performance in the TB task on one hand, and to children's pragmatics understanding in more classic pragmatics tasks on the other hand.

As recursive ToM is claimed to emerge by recursive embedding of representations, recursive ToM itself might, in turn, at least partly be based on a more general ability for recursive thinking. Hence, the set of studies also tested whether recursive forms ToM relate to recursive embedding abilities in general tested in a syntactic recursion task.

In three studies, 232 six- to ten-year-old children were tested. Children were tested in first- to fifth-order ToM tasks. Their recursive ToM abilities were compared to their performance in the TB task, a metaphor and irony pragmatic language task, and a syntactic recursion task.

Results reveal major developmental progress in higher-order ToM beyond first and second-order FB understanding from age six to ten. With increasing age, children came to solve third-order FB tasks, and more rarely, even fourth and fifth-order FB tasks. Results also revealed substantial correlations to recursive embedding abilities in other areas suggesting a parallel development of higher-order recursive thinking. However, performance in recursive ToM tasks was not related to children's metaphor and irony understanding.

Results on the relationship between recursive ToM and recovering TB performance were mixed. Study 1 suggested that children's recursive ToM predict their TB performance as children with higher orders of mental state reasoning were more likely to solve the TB correctly. This relationship, however, could not be replicated in studies 2 and 3.

With regard to research questions (Ib) How recursive meta-representational ToM develops throughout childhood and how it can be measured, and (IIb) whether higher-order ToM development is related to the ability of recursive thinking, project 2 provides refined measures and evidence of subsequent gradual development of recursive meta-representational development in ToM after the four-year revolution and correlated performance in syntactic recursion abilities.

### 7.4. Project 3 – Schidelko, Baumann, Proft, & Rakoczy (2023)

Based on these results of subsequent development of recursive ToM and parallel development in syntactic non-representational recursion abilities, the third project investigated how recursive ToM development is related to other cognitive abilities. As more complex and higher-order forms of mental state reasoning emerge by the recursive embedding of representations, recursive ToM development might be related to other abilities that are also based on the recursive embedding of representations. The argument has been made that advanced Mental Time Travel necessitates representing how past or future events are related to the present state, and therefore depends on some type of meta-representation. This idea was approached by comparing recursive ToM with Mental Time Travel on different orders of meta-representational complexity. Building on the concept of a shared cognitive foundation, we hypothesized a parallel and correlated onset of meta-representational thinking in the domains of ToM and temporal cognition, as well as subsequent development in recursive forms of both (Bieri, 1986; Gautam et al., 2019; Redshaw, 2014).
120 three to eight-year-old children were tested in six tasks that tested for their first, second, and third-order FB ToM understanding as well as for their first, second, and third-order Mental Time Travel abilities. First-order Mental Time Traveling abilities were operationalized as reasoning about future possibilities. Second-order Mental Time Traveling tested for counterfactual thought and third-order Mental Time Traveling tested for anticipated counterfactual reasoning. Results revealed a prolonged, stepwise development of orders in ToM and Mental Time Travel for each domain separately and a moderate consistency between the level of development across domains. However, to a large degree, this effect was caused by a correlation between performance in second-order ToM and second-order Mental Time Traveling. These results of Project 3 extend on evidence from Project 2 on substantial, stepwise developmental progress in meta-representational ToM after the four-year revolution. With regard to the research question (IIa), whether ToM emergence and subsequent development is related to a general understanding of

representations, Project 3 provides evidence of parallel but only partly correlated trajectories of higher-order ToM and Mental Time Travel abilities.

# 8. General Discussion

This dissertation aimed to investigate the emergence and development of the meta-representational ToM. To this end, this dissertation presented ways to systematically test mental and non-mental representational thinking in development and systematically compared patterns of performance in tasks that require (recursive) mental meta-representational thought, (recursive) non-mental representational thought, and no meta-representational tasks. These systematic task comparisons were implemented in three projects. As most of the studies were conducted remotely via video-conferencing systems, an additional preliminary project tested the validity of this new method of data collection.

In the General Discussion of this dissertation, I will interpret the findings in light of domain-general ideas of ToM and will propose directions for future research. In the first part of the General Discussion, I will focus on a general understanding of representations and discuss differences in pragmatic demands in tasks used to test this understanding in different domains. The second part of the General Discussion will discuss the role of Executive Functions in ToM development and the potential underlying cognitive structures of Executive Functions, ToM development, and other representational abilities. Finally, I will give an outlook on future research investigating the relationship between ToM and modal thought in development.

## 8.1. Meta-Representational Thought

### 8.1.1. Joint Emergence of First-Order Meta-Representational Thought

Recent results have questioned the standard picture of the onset of meta-representational ToM around age four: Once children master the FB task, they begin to fail TB control tasks. Different accounts have been proposed to explain this phenomenon. Overestimation accounts, such as Perceptional Access Reasoning (PAR) and the Mental File Card Theory (MFCT), claim that children's performance in the TB Task shows a lack of a genuine ToM in children even beyond age four. Instead, children use a simpler heuristic based on agents' perceptual access to a scenery. Pragmatic accounts, in contrast, assume that the TB task is pragmatically confusing because it poses a trivial academic test question about a rational agent's perspective. The results from project 1 are highly compatible with pragmatic performance limitation accounts on the TB task and with an onset of meta-representational thought around age four. Children only showed difficulties in the trivial academic TB test questions that refer to mental representations of rational agents but not in the trivial academic test questions that refer to a non-mental representation of a signpost. Additionally, results of positive associations of FB and False Sign performance support analogous development in mental and non-mental representational thought (Bowler et al., 2005; Iao & Leekam, 2014; Sabbagh et al., 2006).

Overall, the results help to approach the answers to research questions (Ia) When does meta-representational ToM emerge?, and (IIa) How does Theory of Mind emergence relate to meta-representational thought in other domains?, by providing evidence of a joint developmental transition marking the onset of meta-representation and perspective understanding around age four (Moll et al., 2013; Perner et al., 2003; Perner, Stummer, et al., 2002; Perner & Roessler, 2012). In line with domain-general views, ToM emerges in the course of a general transition to meta-representational thought around age four and pragmatic performance limitation can account for failure in TB tasks caused by three task factors: the (1) academic and (2) highly trivial question that refers to a (3) subjective mental representation.

### 8.1.2. Parallel development of Higher-Order Thought

Building on the results of a unified onset, project 2 and 3 investigated subsequent development of higher-order ToM. Higher-order ToM emerges by recursive embedding of mental representations. Therefore, it was argued that such complex forms of ToM build on higher-order meta-representational thought and general recursive reasoning abilities. Project 2 investigated developmental relations between higher-order ToM and recursive thinking and project 3 approached cognitive foundations of higher-order ToM in higher-order meta-representational thought by comparing complex forms of ToM with meta-representational Mental Time Travel.

The results revealed major, stepwise developmental progress in higher-order Theory of Mind after age four. Similarly, parallel, stepwise development was found in recursive Mental Time Travel abilities. Furthermore, higher-order ToM development was substantially correlated with general recursive reasoning abilities. However, performance in higher-order ToM and recursive Mental Time Travel was only partly associated.

Together, these results help to approach the answers to research questions (Ib) How does higher-order ToM develop?, and (II) How does ToM development relate to (IIa) meta-representational thought in other domains, and (IIb) to general recursive reasoning abilties, by providing evidence of major developmental progress in higher-order Theory of Mind beyond second-order in children between six and ten years. The findings are compatible with domain-general theories of ToM development: To a certain extent, developmental progress in higher-order Theory of Mind base on shared abilities as meta-representational thought and recursive reasoning. This evidence contributes to a more comprehensive understanding of the cognitive abilities involved in ontogenetic Theory of Mind.

#### 8.1.3. Pragmatics: Agency and Intentionality

At the same time, the results of project 1 leave open many fundamental questions. Why are children – with growing socio-cognitive and pragmatic abilities – confused by a trivial academic question that refers to a mental representation but, at the same time, not confused by a maximally similar question about a non-mental representation (the signpost)? Why does a question that pertains to a mental subjective perspective evoke the possibility of a different perspective and, thus, of error, and why does a question that pertains to a non-mental perspective not? One speculation is that we make different kinds of rationality assumptions with regard to the rational agent in the TB task and to external signs in the True Sign task as a consequence of their different forms of Intentionality. This difference may explain the deviating pattern of results in the otherwise parallel tasks. In the following, I will explain what is meant by Intentionality, spell out the idea in more detail, and propose possible ways to test it empirically.

From a philosophy of mind point of view, we see others and ourselves as rational agents and as *ipso facto* radically different from how we perceive and describe inanimate things in the world (as, for example, signposts) (*e.g.*, Rakoczy, 2017). Rational agents are seen as performing rational actions that are based on reasons: conative mental states (as desires) and cognitive mental states (as beliefs). By ascribing desires and beliefs to rational agents, we can explain or predict their rational actions (Davidson, 1963). In the TB task, the protagonist, Maxi, can be seen as an agent who holds beliefs and desires and enough rationality as to what it ought to do given those beliefs and desires. The TB test question ("Where will Maxi search for his chocolate?") asks for such an action prediction of the protagonist Maxi. To answer the question in the TB task, we, therefore, base our action prediction on his desire (that he wants to have the chocolate) and his belief (where he believes the chocolate to be, let's say the green cupboard, and that is, where the chocolate really is and where the child, the experimenter, and Maxi know that it is). So, both the conative and cognitive reasons for his actions are common ground in the whole course of the task. His action can easily be predicted: Maxi will go to the green cupboard.

Given the fact that his beliefs (or in most tasks: knowledge) and desires are common ground during the whole course of the task and the only rational action in general and also in light of his beliefs and desires is to go to the green cupboard, the question arises why the experimenter asks this question. What is the point of asking the question? Accounts of pragmatics commonly presuppose that speech acts have to be of any relevance and that listeners interpret speech acts towards the most relevant interpretation (Grice, 1975; Sperber & Wilson, 1986).

Now, where is this relevance in the TB task when the test question asks for something that is both obvious and common ground? It seems that the experimenter violates the assumption of relevance in asking the test question in the TB task. Children might try to find the relevant point in asking this question (Rakoczy & Oktay-Gür, 2020). As Maxi's desire is usually clearly stated at the end of the storyline (*e.g*., "Maxi returns and wants to have his chocolate now"), children might try to find the relevant point in the cognitive reason giving part, *i.e.,* Maxi's underlying belief (in some studies, the test question indeed directly asks for a belief ascription "What does Maxi believe where the chocolate is?") Such a question seems pragmatically unnatural (Papafragou et al., 2007): We ask for and talk about beliefs only if there is at least some possibility of a divergent perspective or misrepresentation. Asking the TB test question in this situation, therefore, implies that there may be an alternative perspective or misrepresentation involved. Children may thus look for such a possible alternative perspective on the scenario. A divergent perspective on the scenario that critically changes the prediction of action is a FB about the object's location. If the agent held the outdated belief that the chocolate is still in the first location, the blue cupboard, the action prediction would change in that we predict him to go to this location. This way the children might impose relevance on the speech act of asking the TB test question.

Given the consistently high performance in the True Sign task reported in Project 1, it is worth considering why the same level of performance is not observed in the sign task. The True Sign tasks follows a parallel storyline with the only difference that a signpost (and not an agent) ought to represent the object's location. The True Sign test question might therefore be equally unnatural as it asks for a (non-mental) perspective in a scenario in which everything is shared and common ground.

What is different, however, is that the signpost is probably not seen as a rational agent to whom we apply the same rationality assumptions. We might make different kinds of rationality assumptions dependent on the Intentionality of the entity: the rational agent's mental states or the external signpost. What affects this difference in rationality assumptions made for agents' mental states versus representations of external signs?

One core feature of cognitive and conative mental states is that they are directed or about something. An agent holding a belief holds a belief *about something,* believing that something is the case. Likewise, an agent cannot hold a desire without desiring *something*. This feature is called Intentionality[8] (lat.: *intendere* - being directed towards some target or thing). The Intentionality of mental states is intrinsic and original. Mental states are inherently *directed towards something* or *about something* and do not require an external interpretation to become or be used as representations (Searle, 1983; *but see* Dennett, 1994). A belief or a desire simply consists of *Intentional content* in a *psychological mode* (Searle, 1983). For example, in the FB story, Maxi *believes* (the psychological mode or manner of representation) *that the chocolate is the green cupboard* (the Intentional content, in the case of beliefs, propositional content). Intentional content designates its *condition of satisfaction*. Beliefs hold the condition of satisfaction that the content fits reality. A belief is true if it matches reality, but false if it does not. Desires, in contrast, hold the condition of satisfaction that a certain state of the world changes accordingly to the desire. They aim at being fulfilled or realized. The psychological mode designates that these conditions of satisfaction are represented with a particular *direction of fit*. As beliefs aim at being accurate, they have a *mind-to-world direction of fit*. Desires aim at adjusting the world and thus have a *world-to-mind direction of fit* (Searle, 1983).

The Intentionality of artifacts including external signs, in contrast, is neither intrinsic nor original. External signs, just like pictures and sentences, have a *derived* form of Intentionality. Their derived Intentionality implies that external signs represent in a different manner compared to Intentional states, which have an intrinsic form of Intentionality. An interpreter is required who uses the entity as a representation for there to be a representation. In order to get from the mere physical realization (the actual signpost) to the representation, the interpreting mind needs to impose Intentionality on the non-intrinsically Intentional signpost. The mind does this in that it Intentionally confers the conditions of satisfaction of the conveyed psychological state upon the external physical entity. In the True Sign task, the signpost is installed in a way that it displays the object's current location. Its condition of satisfaction is hence that it fits reality and, thus, is intended to have an "external sign-to-world" direction of fit. The recipient needs to recognize this intention that confers meaning to the sign (Searle, 1983).

After pointing out these different kinds of Intentionality of external and internal representations, the question arises: How might the form of Intentionality – original versus derived – influence

[8] Intentionality in this sense is usually spelled with a capital 'I' in order to distinguish it from intentionality as "intending something".

the interpretation of the speech act of the test question? In the TB task, it seems that the experimenter violates the assumption of relevance in asking the test question that pertains to a mental state with intrinsic or original Intentionality. The original Intentionality of the mental state *intrinsically* determines its condition of satisfaction to fit reality and its mind-to-world direction of fit. As the condition of satisfaction is fulfilled in this case as Maxi's belief fits reality, the question seems to lack relevance and seems unnatural. Children might thus try to find the relevant point in asking this question (Rakoczy & Oktay-Gür, 2020). Attributing a FB about the object's location to the agent would fill this lack of relevance. The belief that the object is still in the first location, would change the action prediction in that we expect him to go to the first location and thus leads to the typical error that children predict that the agent will look for the object in the first location.

In the True Sign task, in contrast, the feature of derived Intentionality of the signpost might provide enough relevance to the speech act and thus prevent children from trying to find the relevant point in asking this question. In the True Sign task, the Intentionality of the entity in question is not intrinsic and thus does not *intrinsically* determine its condition of satisfaction to fit reality and its mind-to-world direction of fit. Instead, the child is required to use the signpost as a representation. She needs to impose Intentionality to confer the conditions of satisfaction to fit reality and, to have an "external sign-to-world" direction of fit. This need for imposing Intentionality to the signpost might fill the lack of relevance given in the TB task.

In summary, the idea is that children hold diverging rationality assumptions with regard to the rational agent in the TB task and of external sign in the True Sign task as a consequence of their original versus derived Intentionality. The original Intentionality of the mental state in question in the TB task might convey a lack of relevance to the speech act. The derived Intentionality of the signpost in the True Sign task, in contrast, requires the child to use the sign as a representation and to impose Intentionality on it to confer meaning. These requirements might provide enough relevance to the speech act and thus do not lead the child into pragmatic confusion about the question that leads them to find a possibility for misrepresentation or error in the scenario.

#### 8.1.3.1. Future Directions on Intentionality and Pragmatics of Test Questions

As this is, so far, not more than speculation based on the empirical results of project 1, how could future studies test for this possibility empirically?

First, the entity in the task could be gradually varied regarding their kind of Intentionality. In the comparison of the TB task and the True Sign task, two clearly distinct categories of entities are used. The TB task involves a human protagonist who is perceived as a highly rational agent with intrinsic intentional states. The signpost in the True Sign task, in contrast, is a simple artifact that can be seen to derive its meanings exclusively from its functional role and has no intrinsic meaning that would exist if we did not interpret it (Dennett, 2009). It has been discussed whether more sophisticated artifacts, such as computer or robots possess a kind of Intentionality that is imbued by their creator and can unfold without any direct dependence of their users. These artifacts are thought to have some kind of internal states with a sort of meaning to them that may be unknown to the user. This approach would allow researchers to investigate the extent to which users of such sophisticated artifacts adopt an Intentional stance when explaining the behavior of the artifact (*e.g*., Dennett, 2009). This provides a perspective from which to investigate the gap between questions pertaining to the intrinsic mental states' highly rational human agents and simple artifacts. To this end, the exact same storyline of the change-of-location task could be implemented with a sophisticated artifact, such as robots with graded human features, representing the object's location. The predictions would be the following: more simple artifacts (as the non-directional sign) would be interpreted as systems with derived Intentionality to which children need to confer meanings and, thus, do not perceive a lack of relevance in asking about the system's Intentional state. When increasing the complexity of the system and designing them as more sophisticated artifacts, such as a sign with a motion-detection sensor and a screen or a human-like robot with motion and voice, the systems may be perceived as having agency. If children fail versions of this task with more human-like protagonists than tasks with more artificial artifact-like protagonists, this would support the idea that children's pragmatic interpretation of the test question is influenced by the kind of Intentionality and the rationality assumptions made based on the kind of Intentionality. In this way, future investigations could contribute interesting insights into how children perceive agency and how this interacts with their pragmatics understanding in communication involving agency.

Further insights about how Intentionality influences children's pragmatics understanding might come from a second line of research. If indeed children find trivial academic test questions about a rational agent's TB pragmatically confusing, how general is this phenomenon? (Oktay-Gür and Rakoczy, 2017). Would it hold in similar ways for the ascription of other mental states, for example, in a task in which the test question asks for a trivial desire and the belief is clearly mentioned? Think of a scenario of the following kind: "Maxi is in the kitchen, he is hungry, and his apple is in the blue cupboard just in front of him. Maxi loves apples and likes to eat them when he is hungry and knows that they are good for him. He knows his apple is in the blue cupboard. What does he want to do?" On the one hand, asking this question might be similarly confusing as asking the test question in the original TB task. What is the point of asking about Maxi's desire in this situation in which it is completely obvious what is good and to be done? Following the analysis applied above, children might try to find the relevance of this speech act about an intrinsic Intentional state (the desire). In this case, such a task could lead to similar results of failure as the TB task. On the other hand, the condition of satisfaction and thus, the direction of fit differs of the Intentional state in question differs. Desires hold the condition of satisfaction that a certain state of the world changes accordingly to the desire. They aim at being fulfilled or realized. Desires aim at adjusting the world and thus have a *world-to-mind direction of fit*. Thus, when asking the question about Maxi's desire, the condition is not (yet) fulfilled. The unfulfilled desire thus could provide enough relevance for the speech act so that children are not pragmatically confused about this question compared to a question about a belief its condition of satisfaction is fulfilled as it fits reality. Preliminary evidence for an effect of a highlight of the agent's knowledge state comes from studies that highlighted the agent's visual access to the scene in the classic TB task (Schidelko & Rakoczy, *in preparation*). In this version of the TB, the experimenter highlighted the protagonist's visual access ("The chocolate is now in the blue box. The *protagonist* saw that") before asking for the action prediction ("Where will the agent look for the object?"). Importantly, in this study, children were asked for the action prediction and not, explicitly, for the agent's desire. However, the underlying rationality assumption is that the agent will perform a rational action based on a combination of beliefs and desires. And the preceding highlight ("The agent saw that") clarifies the agent's belief. When children now try to interpret the experimenter's speech act with regard to its relevance, they might interpret the question to ask for the missing piece of information. That might draw children's attention to the action-causing desire (the desire to have the object). And the condition of the desire, in contrast to the belief, is not satisfied as the desire is not (yet) fulfilled. In line with this analysis, preliminary evidence administering this version of the TB task led to success in the TB task in four- to seven-year-old children (Schidelko & Rakoczy, *in preparation*). Future studies investigating whether children show the same difficulties in answering a trivial question about someone's desire will provide interesting insights into the mechanisms and the scope of children's difficulties in the TB task. From a broader perspective, such investigations will contribute to children's perception of agency and rationality and resulting relevance assumptions.

## 8.2. Developmental Determinants of the Emergence of ToM

The results of project 1 support the claim of the onset of ToM in the course of a general transition to meta-representational thought around age four (Perner, 1991; Wellman et al., 2001; Wimmer & Perner, 1983). From a domain-general view of ToM, this raises the question of proximal mechanisms that drive this development of meta-representational thought. Which cognitive capacities drive development that allows us to hold parallel subjective perspectives in mind and to understand representations as mere representations?

Besides some general cognitive, neural, and environmental determinants (Aldrich et al., 2021; Devine & Hughes, 2018; Ebert et al., 2017; Meins et al., 2002, p. 200; Xiao et al., 2019), general executive abilities have been proposed as a main candidate of cognitive capacities ToM development is influenced by or even builds on (Frye et al., 1995). In the following, I will present evidence for the relatedness of ToM and Executive Functions (EF) and discuss the role of EF in the emergence and development of ToM concerning the results presented in this dissertation.

### 8.2.1. Executive Functions and the Emergence of ToM

Children's EF abilities have been suggested to play a crucial role in the development of ToM. EF is used as an umbrella term for a multidimensional construct of cognitive skills broadly related to mechanisms of cognitive control, attention, and flexibility (Anderson, 2002; Carlson, 2005). EF is usually conceptualized and behaviorally measured as three skills – inhibitory control, working memory, and cognitive flexibility (Garon et al., 2008; see also Miyake et al., 2000). These skills, in combination, allow individuals to control their thoughts and behavior (Oh & Lewis, 2008). On a behavioral level, evidence shows that children significantly improve in EF tasks during preschool age. Across a wide range of EF tasks, performance is consistently related (Miyake et al., 2000; Wiebe et al., 2011).

Crucially, not only is performance in EF related among EF tasks, but it also shows close developmental links to ToM in children (Carlson & Moses, 2001; Perner & Lang, 1999), across the lifespan (Qureshi et al., 2010; Rakoczy et al., 2012), and in patients with neuropsychological impairments (*e.g.*, Samson et al., 2005). Correlational studies have shown cross-sectional associations between the two abilities (Carlson et al., 2002; Hughes, 1998), and longitudinal and training studies with robust confound controls have allowed for a causal interpretation of the interrelatedness of ToM and EF. Longitudinal findings show that EF skills in early development predict ToM performance in FB tasks later in development (Carlson et al., 2004; Marcovitch et al., 2015). Importantly, this association is asymmetric: early ToM does not predict later EF (for a meta-analysis, *see* Devine & Hughes, 2014). In training studies, EF skills mediate how fast and to what extent children benefit from ToM training experience over time (Benson et al., 2013). Overall, there is strong evidence for an association between EF and (first-order) ToM in development, with the substantial influence of EF on the emergence of ToM (Devine & Hughes, 2014). However, most research has focused on the relationship between EF and the emergence of first-order ToM. What role does EF play in the subsequent development of the higher-order ToM?

### 8.2.2. Executive Function and the Development of Higher-Order Theory of Mind

Before discussing the role of EF in subsequent development of ToM, it is important to consider the developmental relationship between the emergence and development of ToM to evaluate to what extent the emergence of first-order ToM and development of more complex forms of ToM reflect continuous development of the same ability. Therefore, I will shortly present some evidence regarding the developmental (dis)continuities between first-order and higher-order ToM before discussing the role of EF in the subsequent development of ToM.

Concerning the developmental relationship between the emergence and development of ToM, evidence from longitudinal studies has shown some developmental continuity between first- and second-order FB understanding. Studies that tested for children's first-order FB understanding at an earlier time point and for their second-order FB understanding at a later time point found small to moderate correlations between the two measures (Hughes, 1998; Lockl & Schneider, 2007; Peskin et al., 2016; *but see* Guajardo & Cartwright, 2016). However, little is known about the developmental relations of first-order ToM to higher (above second-) order ToM. One cross-sectional study, with five to eight-year-old children, compared children's first-order FB understanding with diverse AToM tasks, including third and fourth-order FB understanding as well as the Strange Stories task and the Reading Minds in the Eyes test. Children's first-order FB understanding correlated weakly with their social reasoning AToM abilities after controlling for their inhibition and language skills (Osterhaus & Koerber, 2021). It is thus still an open question how the emergence of first-order ToM and the ability of higher-order mental state ascription relate to each other in development. With regard to the broader concept of AToM abilities, earlier first-order FB performance showed weak associations with later performance in diverse AToM tasks, such as the Strange Stories task in longitudinal studies (Devine et al., 2016; Ebert, 2020; Osterhaus et al., 2022). In summary, preliminary evidence suggests some developmental continuity between first- and higher-order ToM and weaker associations between first-order ToM and AToM. This is compatible with the ideas that AToM incorporates a broader set of social skills while higher-order ToM reflects a continuing development based on first-order ToM emergence.

Regarding the role of EF in the subsequent development of ToM, most work concerns the broader construct of AToM but does not investigate or report the role of EF for the specific ability of higher-order ToM. For the broad construct of AToM, evidence shows concurrent but no longitudinal associations between ToM (tested with, among others, the Strange Stories task) and EF with a mean testing interval of more than four years in six- to ten-year-olds (Devine et al., 2016). When the lag between earlier and later testing is shorter (twelve-months period), longitudinal analyses revealed an asymmetric developmental relation between AToM (tested with the Strange Stories tasks and a Silent Film task) and working memory. Early working memory predicted later ToM but not vice versa suggesting a specific role of working memory in the ongoing development of ToM in nine- to ten-year-old children (Lecce et al., 2017).

Regarding the narrower construct of higher-order ToM, examinations were mainly limited to second-order ToM and did not test for higher-order ToM beyond second-order recursion. The comparisons of EF and second-order reasoning showed cross-sectional correlations, and most of the correlations remained significant when controlled for age or IQ (Peloquin et al., 2022; Perner, Kain, et al., 2002). One study conducted with adolescents and adults found associations between higher-order ToM tasks and working memory (Valle et al., 2015).

In summary, evidence suggests that the relationship of EF to first-order ToM emergence differs from its relation to subsequent development in ToM. Although EF showed a clear longitudinal effect on later first-order ToM earlier, evidence for such longitudinal associations between EF and the development of higher-order ToM is rather limited. Instead, evidence points to cross-sectional associations between higher-order and advanced ToM and EF. Due to the broad construct of AToM and the variety of measures used, this relationship is less clear.

### 8.2.3. Accounts on the Relationship of Executive Function and Theory of Mind

Different proposals have been made concerning the nature of the relationship between ToM (emergence) and EF. Some views claim that ToM emergence depends on EF. Other views claim that both ToM and EF depend on a third factor[9]. These accounts are most compatible with the current state of evidence on the relation between EF and ToM and will, thus, be introduced.

*Emergence* accounts focus on the claim that EF is part of the cognitive foundation that enables the flexible employment of mental states, including the coordination and confrontation of mental states, as well as multiple perspectives more generally (Fizke et al., 2014; Perner et al., 2003; Perner, Kain, et al., 2002; Rakoczy, 2010, 2022). Children's conceptual deficit before ToM emergence, hence, lies in flexibly coordinating and employing mental states. The idea receives support from studies that show a specific relationship between EF and ToM tasks that require the coordination of multiple, incompatible perspectives. In these studies, EF was related to children's coordination of different kinds of incompatible mental perspectives, for example, conflicting desires and beliefs (Fizke et al., 2014, Rakoczy 2010). Interestingly, this relationship is not limited to mental perspective problems; EF tasks also correlate with tasks that require an understanding of non-mental perspective problems, such as false signs (Sabbagh et al., 2006). The studies reveal that the so-called 'conflict inhibition' EF tasks specifically account for these correlations. Conflict inhibition tasks, which ask children to comply with the commands of a nice bear but neglect the commands of a mean dragon (*e.g.*, Carlson & Moses, 2001; Carlson et al., 2002; Hughes, 1998; *for an overview*, see Moses et al., 2005), incorporate both inhibition and working memory demands (Fizke et al., 2014, Rakoczy, 2022). Based on this evidence, it has been argued that EF at least partly provides the cognitive foundations to flexibly coordinate perspectives and is thus fundamental for the ability to hold meta-representations (Fizke et al., 2014; Leekam et al., 2008; Perner et al., 2003; Perner & Roessler, 2012; Rakoczy, 2010, 2022; Sabbagh et al., 2006, 2006).

Some Emergence accounts suggest that EF capacities themselves are the cognitive foundation of ToM and perspective understanding. The accounts mainly claim that EF plays a scaffolding role: EF is necessary for children to develop an understanding of mental state concepts (*e.g.,* Carlson & Moses, 2001). In contrast, other accounts[10] propose an underlying mechanism through which EF enables the emergence of ToM. According to such views, both ToM and EF

[9] Earlier views which based on merely correlational findings also addressed the possibility that EF depends on Theory of Mind (Kloo & Perner, 2003, 2005). However, evidence of an asymmetrical relationship of EF on ToM from training and longitudinal studies made these theories rather unlikely. Other accounts, called *Expression accounts* claimed that Theory of Mind is present from early childhood but young children cannot yet express it due to their limited EF (Fodor, 1992; Mitchell, 1996; Russell, 1992). Certain development in EF is therefore required in development for children to express their ToM. Expression accounts are, however, not compatible with findings that show that children's ToM performance does not increase significantly when EF task demands are reduced in ToM tasks (*see* Wellman et al., 2001). Thus, these accounts will be disregarded here (*for overviews, see* Moses & Tahiroglu, 2010; Perner & Lang, 1999).

[10] These accounts are also referred to as *Competence* Accounts (Apperly et al., 2009).

consequently share common cognitive underpinnings. Therefore, they are particularly important when discussing the results on the relation of ToM and (IIa) a general understanding of representations, and (IIb) general recursive reasoning abilities.

Three main accounts proposing underlying cognitive structures will be introduced: recursive reasoning (*e.g.*, Hauser et al., 2002), complex relational reasoning (Andrews et al., 2003; Halford et al., 1998), and hierarchical thinking (Frye et al., 1995). This section will introduce the main ideas of these accounts to discuss which account best explains the results of this dissertation, which suggest protracted development in recursive forms of ToM and parallel, partly correlated development in other cognitive abilities, in light of domain-general views on ToM emergence and development.

### 8.2.4. Underlying Cognitive Structures

First, one group of accounts proposes that both EF and ToM are based on the computational capacity for recursive operations (*e.g.,* Hauser et al., 2002). Recursion is generally referred to as the computational capacity that facilitates us to embed elements within elements of the same kind (Pinker & Jackendoff, 2005). It has been argued that first-order FB understanding requires meta-representational thought which reflects a recursive operation. Holding a meta-representation requires understanding and representing a representation as representation. In other words, a representation, for instance, the agent's FB, has to be embedded in its own representation of the world. That matches the definition of embedding elements (someone's mental representation) within elements of the same kind (their own representation). Similarly, the False Sign task requires the embedding of the representation of the sign into its own representation of the world. From a developmental point of view, it could be argued that children first become able to compute first-order recursive operations and after a developmental delay become able to process second-order and higher-order recursive operations. Based on this idea, it could be argued that children show parallel and interrelated developmental progress in tasks that require the same level of recursive embedding. In general, the results of the project in this dissertation converge with this idea. Children's performance in project 1 on first-order FB and False Sign tasks is correlated while performance in analog non-representational tasks varies.

Second, the Cognitive Complexity and Control theory (*e.g.*, Zelazo & Frye, 1998) logically analyses cognitive processes in terms of hierarchically embedded conditionals, so-called "if-if-then" rules. More complex rule-following abilities (*e.g.*, 'IF [IF a THEN b] THEN [IF c THEN d]') would enable both EF and meta-representational ToM. The theory is based on and is best explained by the common behavioral measure to test EF in development. This is the Dimensional Change Card Sort (DCCS) task (Zelazo et al. 2003; Frye et al., 1995). In this task, participants are given two target cards with pictures on them and are asked to sort the test cards into the target cards. The cards vary in two dimensions, for example, shape and color. First, the participants are asked to sort them, for example, according to their shape (the shape game: squares to squares and triangles to triangles regardless of the shape's color). Next, they are asked to sort them in the other dimension, *i.e.*, according to their color (the color game: red to red cards and blue to blue cards regardless of their shape). The DCCS presumably requires all three skills related to EF: Inhibitory control to suppress the preceding response to sort cards according to the first rule, working memory to hold the current rule in mind, and cognitive flexibility to adapt to the new sorting rule. The logical analysis of this task according to the Cognitive Complexity and Control theory (*see* Figure 2) is the following: *if* we play the shape game (game 1), and *if* the card shows a red square (antecedent 1), *then* you place it with the green square target (consequent 1), but if the card shows a green triangle (antecedent 2), *then* you place it with the red triangle target (consequent 2). In the second game (the color game), the contingencies are reversed, so that antecedent 1 (the red square) maps to consequent 2 (the red triangle target) (Frye et al., 1995; *see also* Perner & Lang, 1999).

The logical analysis of the classic change-of-location FB tasks requires following similarly complex rules: *if* asked for self-perspective (perspective 1), and *if* asked for the chocolate (antecedent 1), *then* the second (true) location (consequent 1), and *if* asked for story protagonist's perspective (perspective 2), and *if* asked for the chocolate (antecedent 1), *then* the first (outdated) location (consequent 1). The first "if" (for the condition, *i.e.* self or story protagonist) allows switching answers flexibly according to the perspective mentioned in the test question (Frye et al., 1995; Zelazo & Frye, 1998). In both conditions, the second consequence may be the other location, however, there are no obvious second antecedents involved in the FB task as in the analysis of the DCCD task. Thus, the FB task can be solved by embedding only one rule under each setting condition while EF tasks such as the DCCS tasks require embedding two rules (Perner & Lang, 1999; Perner et al., 2002). Importantly, the Cognitive Complexity and Control theory defines complexity in terms of hierarchical levels. As in both cases, a higher-order rule needs to be applied. It is suggested that the complexity of rules is similar for FB tasks and the DCCS task. A higher-order rule is required whenever two antecedent conditions (if-if-then) must be processed in order to select which lower-order rule applies (Zelazo, 2004; Zelazo & Frye, 1998). Consequently, becoming able to form and follow rules of this complexity (if-if-then rules) thus leads to success in both EF tasks (as the DCCS task) and meta-representational ToM tasks (Zelazo & Frye, 1998; but *see* Perner & Lang, 1999).

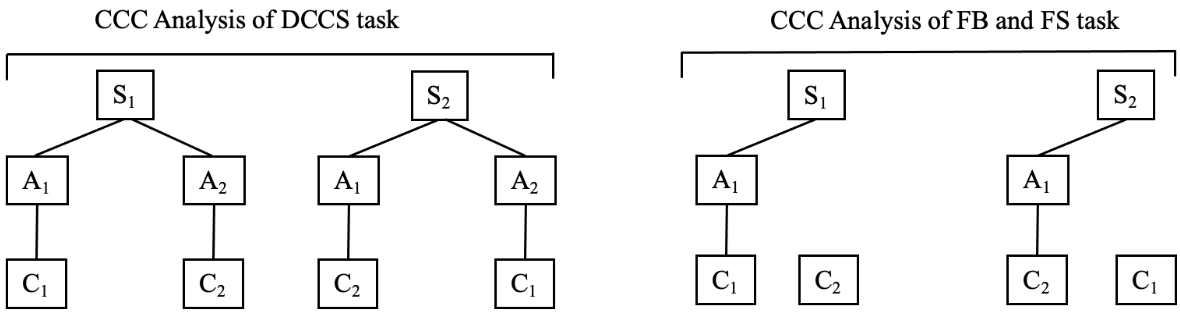


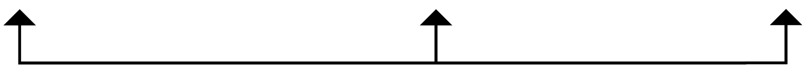

**Figure 2**. Schematic Visualization of higher-order rules based on the Cognitive Complexity and Control (CCC) Theory for the DCCS task (left) and the False Belief (FB) and False Sign (FS) task (right) (adapted from Andrews et al., 2003). *S* refers to the setting condition, *A* refers to the sorting-card attribute, and *C* refers to the sorting category.

The results of project 1 of this dissertation are in line with this idea. According to the Cognitive Complexity and Control theory, performance in tasks that require higher-order rules should be associated, while performance in tasks that can be solved without higher-order rules diverges. The results of project 1 show that children's performance in the FB task and the False Sign task is substantially correlated, whereas performance in the parallel TB task and True Sign task is dissociated. It can be argued that the False Sign task, similar to the FB task and the DCCS task, requires a similar logical rule structure: two antecedent conditions need to be considered in order to select the applying lower-order rule, such as *if* asked for self-perspective (perspective 1), and *if* asked for the vehicle (antecedent 1), *then* the second (true) location (consequent 1), and *if* asked for signpost's perspective (perspective 2), and *if* asked for the vehicle (antecedent 1), *then* the first (outdated) location (consequent 1). The TB and the True Sign task, in contrast, do not offer two antecedent conditions as the self-perspective and the other perspective (of the signpost or the story protagonist) align (*see* Figure 2).

Third, the general idea of the Relation Complexity Theory is that cognitive processes or representations can be defined in terms of their structural complexity development (Andrews et al., 2003; Halford et al., 1998). The number of variables that are related to each other constitutes the *dimensionality* of a relation of a given cognitive representation or process. The general information processing capacity is determined by the complex relations that can be processed in parallel. In that way, the information processing factor influences a wide range of cognitive processes. Depending on the number of variables involved in a certain task, children can or cannot yet represent the relational structures. It has been claimed that the ability to process relations of increasing complexity gradually develops during childhood (Andrews, et al., 2003; Halford et al., 1998). In the typical EF DCCS task, three variables need to be related: the setting condition *S* (shape game or color game), the sorting-card attribute *A,* and the sorting category *C* (*see* Figure 3). Similarly, the FB task requires processing a ternary relation between three variables: the environmental cue (the object location), the setting condition (movement seen or not by the story protagonist), and the person's representation (represented object location, for example, believes object in location 1/ believes object in location 2) (Andrews et al., 2003). The ternary relation allows for representing that an agent's representation of an object's location depends on the interaction between the object's location (*i.e.*, the environmental cue) and the access to the movement (*i.e.*, the setting condition) (Andrews et al., 2003).

According to the Relational Complexity theory, performance in tasks that require relating the same number of variables should be associated, while performance in tasks that require relating more or fewer variables diverges. The False Sign task, like the FB task, also requires relating three variables: the environmental cue (the vehicle's location), the setting condition (sign post changed/ not changed with movement), and the sign post's representation (represents the vehicle at location 1 or 2) (*see* Figure 3). Children's associated performance in the FB task and the False Sign task in project 1 converge with this idea of associated performance in a task that requires representing the same level of dimensionality made by the Relational Complexity theory.

| Task | Environmental cue | Setting condition | Representation |
|---|---|---|---|
| FB | Object Location | Movement seen/ not seen | Represented object location |
| | *e.g., in location 1* | | *Believes object is in location 1/ believes object is in location 2* |
| False Sign | Object Location | Sign changed/ not changed | Represented object location |
| | *e.g., in location 1* | | *Shows object being in location 1/ Shows object being in location 2* |



Ternary relation

**Figure 3**. Schematic application of the Relational Complexity Theory to the FB and False Sign task (*adapted from* Andrews et al., 2003).

#### 8.2.4.1. Underlying Cognitive Structures of Higher-Order Meta-Representation

So far, these underlying cognitive structures have mostly been discussed concerning first-order ToM emergence. In light of evidence of continuing development in ToM the question arises whether they also underlie developmental progress in higher-order ToM? Can they account for the protracted, stepwise development of higher-order forms of ToM and other forms of meta-representational thinking shown in projects 2 and 3 of this dissertation?

On a theoretical level, it has been claimed that second-order ToM is not simply a ToM ability of increased complexity, as it is reflective of a qualitative transformation of the underlying thought system – a *conceptual change* in ToM development (Perner, 1988, *see also* Miller, 2009). At least two (related) cognitive advances after first-order FB understanding have been discussed[11]. First, second - and higher-order ToM but not first-order ToM requires the insight that mental states cannot only be directed to the physical world but also to other minds. The propositional content of beliefs or desires can also concern other mental states: what people think about their own or other people's thoughts (second-order beliefs) (Perner & Wimmer, 1985). Second, as a consequence, second and higher-order ToM requires an understanding of the recursive nature of mental states: propositions can be recursively embedded within each other. This possibility for recursion of mental states needs to be understood to be then applied (Perner, 1988).

Can the discussed underlying cognitive structures account for this conceptual change in subsequent ToM development? In general, the main idea of all three proposed structures would be the same: the increased complexity of the higher-order ToM tasks due to increasing levels of recursive embedding (Braüner et al., 2016; Hauser et al., 2002; Pinker & Jackendoff, 2005; Polyanskaya et al., 2022), increasing number of variables involved (Andrews et al., 2003; Halford et al., 1998), or the increasing hierarchical complexity (Frye et al., 1995) account for the developmental delay after first-order ToM emergence. With developmental progress in the respective cognitive mechanism (recursive embedding, complexity of relations, hierarchical complexity), children become able to solve ToM tasks beyond first-order. While all three theories can account for the protracted development, the recursive embedding approach offers the possibility to integrate the idea of a conceptual change between first-order and higher-order ToM. Therefore, this idea will now be presented and discussed in light of the results of this dissertation.

The idea is the following: The proposed conceptual change from first- to higher-order ToM understanding reflects an underlying change in recursive reasoning from merely using recursive operations to understanding and applying recursive concepts. Children have to *process* a recursive operation for the first-order FB task but have to *understand and apply the concept of recursion* to their concept of mental states only for second-order and higher-order ToM. This understanding of the recursive principle itself and the possibility of its application to mental states might reflect the conceptual change after first-order ToM emergence in children and account for the developmental delay of higher-order ToM (Braüner et al., 2016; Perner, 1988; Polyanskaya et al., 2022). In line with this idea, projects 2 and 3 show developmental delay between tasks that require recursive operations and tasks that require understanding to apply

[11] For other accounts suggesting that the developmental delay of second-order ToM reduced to tasks factors, such as more complex story scenarios and longer test questions, *see e.g.*, Sullivan et al., 1994; Lockl & Schneider.

the concept of recursion. Additionally, performance in tasks requiring recursive operations in different domains was correlated. Despite substantial differences in task features (*e.g.*, length and format of test questions), performance on higher-order ToM tasks and syntactic recursion tasks showed correlated performance even when controlled for children's age. Also, performance on second-order ToM and Mental Time Travel was correlated when controlled for children's age. These results are compatible with the idea of underlying development in the domain of recursive reasoning. Yet, this approach entails the claim that children as soon as they can apply the recursive principle on embedding representations (*i.e.*, solve tasks including second-order embeddings) can process any level of recursive embedding. Contrasting this claim, children showed a developmental delay of success in higher-order ToM and Mental Time Travel tasks after second-order tasks. This leaves several questions open for future research.

First, what happens after the change occurred? Children around six years who solve second-order problems are seen to understand the recursive principle and ought to apply it to meta-representations. Yet, they solve higher-order problems (such as third- and fourth-order tasks) only later during middle childhood. What cognitive limitations prevent them from applying the principle practically infinitely? Potentially, task features other than the level of complexity (such as the length of the storyline and unnecessary information) might cause second-order FB passers' failure in higher-order ToM tasks. To address this hypothesis, task demands of higher-order ToM tasks should be reduced to a minimum. Additionally, higher-order ToM tasks should provide memory aids such as second-order FB tasks (Sullivan et al., 1994). Evidence of a more similar performance in such modified second- and higher-order tasks would provide the first evidence for the influence of task demands on the protracted performance beyond second-order reasoning (Miller, 2009).

Secondly, in relation to the first question, how does an understanding of recursive mental states relate to other forms of recursion and to other complexity analyses? Does recursive reasoning ultimately reflect rule-based reasoning on increasing levels of hierarchical complexity (Cognitive Complexity and Control Theory) or relational complexity (Relational Complexity Theory)? Empirically, this can be tested with systematic comparisons of tasks that can contrast the theories. For example, the Relational Complexity Theory and Cognitive Complexity and Control theory can be contrasted in different versions of the DCCS cart sorting tasks. The original task requires switching between the dimensions and thus requires reasoning based on higher-order rules (Cognitive Complexity and Control Theory). An alternative version of the task in which both dimensions are sorted simultaneously in one metric, in contrast, does not require such a higher-order rule. The Relational Complexity Theory, however, makes no such difference in the prediction for the two versions (Müller et al., 2005). Future versions of the DCCS tasks might systematically vary these features. Additionally, new versions with increased underlying complexity and dimensionality should be invented to compare performance to higher-order ToM and other reasoning abilities, such as Mental Time Travel.

Third, from a domain-general perspective on development in ToM, another question is how widely shared these underlying cognitive structures are. How do the development of ToM and the understanding of (recursive) meta-representational thought, more generally, relate to other forms of complex cognition? Evidence from other lines of research suggests that children show major development in other cognitive abilities around the time they transition to meta-representational thought. For example, children become able to solve tasks that require modal thought. In these tasks, children need to reason based on concepts of *possibility* and *necessity* (*for an overview, see* Leahy & Carey, 2020). In prototypical tasks, children are faced with uncertain situations involving multiple possibilities. To successfully reason about these possibilities, the possibilities need to be marked with modal operators as *merely possible*[12] (Leahy & Carey, 2020). In other words, children have to understand possibilities *as possibilities.* To illustrate this, let's consider the following example: in front of you are three upside-down cups, a green cup, and two red cups and you observe me putting one sticker in the green cup and another sticker in one of the two red cups but you cannot see in which of the two. Now, you are asked to choose one of the three cups to win its content. Which cup do you pick? (Mody & Carey, 2016). Participants tested in these so-called three-cup tasks are required to identify the uncertain options (two red cups) as merely possible and choose the certain option (the green cup) over these two merely possible options. Evidence points to a transition from absence to the presence of this capacity to reason about possibilities in children around age four (*for a review*, *see* Leahy & Carey, 2020). The age of onset of the ability to represent possibilities *as* possibilities raise the question how meta-representational reasoning and modal reasoning are related.

Does reasoning about possibilities base on some form of meta-representational thought – as it requires to represent different possible worlds in relation to each other – and therefore, share cognitive foundations with ToM? Or, does ToM require to represent different possibilities and is thus, based on modal reasoning about perspectives? Both theoretical analyses of required processes and empirical comparisons of children's developing performance in ToM and modal

[12] Because of this structure, some of these tasks (i.e., the subgroup of tasks that involve future possibilities), are also employed to test for children's ability for future oriented Mental Time Travel, see section 3.1.1. How to Meausre Mental Time Travel.

reasoning tasks are needed to address this question of shared underlying foundations to shed light on the debate of domain-general and domain-specific mechanisms of ToM.

## 9. Conclusion

The projects conducted in this dissertation provide evidence for the emergence of meta-representational ToM around age four and related development in meta-representational thought involving non-mental representations. Furthermore, the results show protracted, stepwise development in higher-order forms of ToM after the emergence around age four. Development in higher-order ToM is correlated with children's developing general recursion abilities and shows parallel, but only partly correlated trajectories to stepwise development in Mental Time Travel abilities of increasing complexity.

The findings of this dissertation contribute to a comprehensive understanding of how ToM develops ontogenetically. In accordance with domain-general views on ToM, the results of this dissertation suggest that ToM relies (to a certain extent) on shared, domain-general processes. Future studies need to investigate to what extent performance in recursive ToM and other representational abilities is actually associated with development to clarify the extent to which underlying abilities are shared with other cognitive abilities and whether and to which extents ToM requires additional cognitive abilities specialized to the purpose of ToM computations.

# 10. References

Aldrich, N. J., Chen, J., & Alfieri, L. (2021). Evaluating associations between parental mind-mindedness and children's developmental capacities through meta-analysis. *Developmental Review*, *60*, 100946. https://doi.org/10.1016/j.dr.2021.100946

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology*, *8*(2), 71–82.

Andrews, G., Halford, G. S., Bunch, K. M., Bowden, D., & Jones, T. (2003). Theory of Mind and Relational Complexity. *Child Development*, *74*(5), 1476–1499. https://doi.org/10.1111/1467-8624.00618

Angeleri, R., & Airenti, G. (2014). The development of joke and irony understanding: A study with 3- to 6-year-old children. *Canadian Journal of Experimental Psychology*, *68*(2), 133–146. https://doi.org/10.1037/cep0000011

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825–839. https://doi.org/10.1080/17470218.2012.676055

Astington, J. W. (1993). *The Child's Discovery of the Mind*. Harvard University Press.

Banerjee, R., Watling, D., & Caputi, M. (2011). Peer Relations and the Understanding of Faux Pas: Longitudinal Evidence for Bidirectional Associations: Peer Relations Faux Pas. *Child Development*, *82*(6), 1887–1905. https://doi.org/10.1111/j.1467-8624.2011.01669.x

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407–418. https://doi.org/10.1023/A:1023035012436

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the

Mind in the Eyes” Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Beck, S. R., & Crilly, M. (2009). Is understanding regret dependent on developments in counterfactual thinking? *The British Journal of Developmental Psychology*, *27*(Pt 2), 505–510. https://doi.org/10.1348/026151008X401697

Beck, S. R., & Guthrie, C. (2011). Almost thinking counterfactually: Children’s understanding of close counterfactuals. *Child Development*, *82*(4), 1189–1198.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children’s Thinking about Counterfactuals and Future Hypothecials as Possibilities. *Child Development*, *77*(2), 413–426.

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

Bieri, P. (1986). Zeiterfahrung und Personalität. *Zeit, Natur Und Mensch. Beiträge von Wissenschaftlern Zum Thema „Zeit"*, 261–281.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31. https://doi.org/10.1016/S0010-0277(00)00096-2

Bosacki, S. (2000). Theory of mind and self-concept in preadolescents: Links with gender and language. *Journal of Educational Psychology*, *92*(4).

Bowler, D. M., Briskman, J., Gurvidi, N., & Fornells-Ambrojo, M. (2005). Understanding the Mind or Predicting Signal-Dependent Action? Performance of Children With and Without Autism on Analogues of the False-Belief Task. *Journal of Cognition and Development*, *6*(2), 259–283. https://doi.org/10.1207/s15327647jcd0602_5

Braüner, T., Blackburn, P., & Polyanskaya, I. (2016). Recursive belief manipulation and second-order false-beliefs. *Logic, Language, Information, and Computation: 23rd International Workshop*, 125–144.

Brooks, R., & Meltzoff, A. N. (2002). The importance of eyes: How infants interpret adult looking behavior. , 38(6),. *Developmental Psychology*, *38*(6), 958–966.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49–57. https://doi.org/10.1016/j.tics.2006.11.004

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

Byrne, R. M. J. (2016). Counterfactual Thought. *Annual Review of Psychology*, *67*(1), 135–157. https://doi.org/10.1146/annurev-psych-122414-033249

Call, J., & Tomasello, M. (1999). A Nonverbal False Belief Task: The Performance of Children and Great Apes. *Child Development*, *70*(2), 381–395. https://doi.org/10.1111/1467-8624.00028

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., Tapanya, S., & Singh, S. (2005). Synchrony in the Onset of Mental-State Reasoning: Evidence From Five Cultures. *Psychological Science*, *16*(5), 378–384. https://doi.org/10.1111/j.0956-7976.2005.01544.x

Carlson, S. M. (2005). Developmentally Sensitive Measures of Executive Function in Preschool Children. *Developmental Neuropsychology*, *28*(2), 595–616. https://doi.org/10.1207/s15326942dn2802_3

Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive Function and Theory of Mind: Stability and Prediction From Ages 2 to 3. *Developmental Psychology*, *40*(6), 1105–1122. https://doi.org/10.1037/0012-1649.40.6.1105

Carlson, S. M., & Moses, L. J. (2001). Individual Differences in Inhibitory Control and Children's Theory of Mind. *Child Development*, *72*(4), 1032–1053. https://doi.org/10.1111/1467-8624.00333

Carlson, S., Moses, L., & Breton, C. (2002). How Specific is the Relation between Executive Function and Theory of Mind? Contributions of Inhibitory Control and Working Memory. *Infant and Child Development*. https://doi.org/10.1002/icd

Carruthers, P. (2013). Mindreading in Infancy: Mindreading in Infancy. *Mind & Language*, *28*(2), 141–172. https://doi.org/10.1111/mila.12014

Cheke, L. G., & Clayton, N. S. (2012). Eurasian jays ( *Garrulus glandarius* ) overcome their current desires to anticipate two distinct future needs and plan for them appropriately. *Biology Letters*, *8*(2), 171–175. https://doi.org/10.1098/rsbl.2011.0909

Correia, S. P. C., Dickinson, A., & Clayton, N. S. (2007). Western Scrub-Jays Anticipate Future Needs Independently of Their Current Motivational State. *Current Biology*, *17*(10), 856–861. https://doi.org/10.1016/j.cub.2007.03.063

Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of 'pure reason' in infancy. *Cognition*, *72*(3), 237–267. https://doi.org/10.1016/S0010-0277(99)00039-6

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy*, *60*(23), 685. https://doi.org/10.2307/2023177

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

Del Sette, P., Bambini, V., Bischetti, L., & Lecce, S. (2020). Longitudinal associations between theory of mind and metaphor understanding during middle childhood. *Cognitive Development*, *56*, 100958. https://doi.org/10.1016/j.cogdev.2020.100958

Dennett, D. (1978). Consciousness in nonhuman species: Open Peer Commentary and Authors' Responses. *The Behavioral and Brain Sciences*, *4*(May 2016), 568–570. https://doi.org/10.1017/S0140525X00027862

Dennett, D. (2009). Intentional Systems Theory. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind* (1st ed., pp. 339–350). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199262618.003.0020

Dennett, D. C. (1994). The Myth of Original Intentionality. In *Thinking Computers and Virtual Persons* (pp. 91–107). Elsevier. https://doi.org/10.1016/B978-0-12-215495-9.50009-0

Devine, R. T. (2021). Individual differences in theory of mind in middle childhood and adolescence. In *Theory of Mind in Middle Childhood and Adolescence*. Routledgebe.

Devine, R. T., & Hughes, C. (2013). Silent Films and Strange Stories: Theory of Mind, Gender, and Social Experiences in Middle Childhood. *Child Development*, *84*(3), 989–1003. https://doi.org/10.1111/cdev.12017

Devine, R. T., & Hughes, C. (2014). Relations Between False Belief Understanding and Executive Function in Early Childhood: A Meta-Analysis. *Child Development*, n/a-n/a. https://doi.org/10.1111/cdev.12237

Devine, R. T., & Hughes, C. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the Silent Films and Strange Stories tasks. *Journal of Experimental Child Psychology*, *149*, 23–40. https://doi.org/10.1016/j.jecp.2015.07.011

Devine, R. T., & Hughes, C. (2018). Family Correlates of False Belief Understanding in Early Childhood: A Meta-Analysis. *Child Development*, *89*(3), 971–987. https://doi.org/10.1111/cdev.12682

Devine, R. T., & Lecce, S. (2021). *Theory of Mind in Middle Childhoof and Adolescence: Integrating Multiple Perspectives* (1st ed.). Routledge.

Devine, R. T., White, N., Ensor, R., & Hughes, C. (2016). Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental Psychology*, *52*(5), 758–771. https://doi.org/10.1037/dev0000105

Doherty, M. J. (2009). *Theory of mind: How children understand others' thoughts and feelings*. Psychology Press.

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

Dryll, E. (2009). Changes in Metaphor Comprehension in Children. *Polish Psychological Bulletin*, *40*(4), 204–212. https://doi.org/10.2478/s10059-009-0015-1

Ebert, S. (2020). Theory of mind, language, and reading: Developmental relations from early childhood to early adolescence. *Journal of Experimental Child Psychology*, *191*, 104739. https://doi.org/10.1016/j.jecp.2019.104739

Ebert, S., Peterson, C., Slaughter, V., & Weinert, S. (2017). Links among parents' mental state language, family socioeconomic status, and preschoolers' theory of mind development. *Cognitive Development*, *44*, 32–48. https://doi.org/10.1016/j.cogdev.2017.08.005

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

Fabricius, W. V., Gonzales, C. R., Pesch, A., Weimer, A. A., Pugliese, J., Carroll, K., Bolnick, R. R., Kupfer, A. S., Eisenberg, N., & Spinrad, T. L. (2021). Perceptual Access Reasoning (PAR) in Developing a Representational Theory of Mind. *Monographs of the Society for Research in Child Development*, *86*(3), 7–154. https://doi.org/10.1111/mono.12432

Fabricius, W. V., & Imbens-Bailey, A. (2000). False beliefs about false beliefs. In *Children's* *reasoning and the mind* (pp. 281–294). Psychology Press.

Fabricius, W. V., & Khalil, S. L. (2003). False beliefs or false positives? Limits on children's understanding of mental representation. *Journal of Cognition and Development*, *4*(3), 239–262. https://doi.org/10.1207/S15327647JCD0403_01

Filippova, E., & Astington, J. W. (2008). Further development in social reasoning revealed in discourse irony understanding. *Child Development*, *79*(1), 126–138. https://doi.org/10.1111/j.1467-8624.2007.01115.x

Fink, E. (2021). Friendships and theory of mind in middle childhood and adolescence. In R. T. Devine & S. Lecce (Eds.), *Theory of Mind in Middle Childhood and Adolescence* (1st ed., pp. 122–141). Routledge. https://doi.org/10.4324/9780429326899-9

Fizke, E., Barthel, D., Peters, T., & Rakoczy, H. (2014). Executive function plays a role in co-ordinating different perspectives, particularly when one's own perspective is involved. *Cognition*, *130*(3), 315–334. https://doi.org/10.1016/j.cognition.2013.11.017

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young Children's Knowledge About Visual Perception: Further Evidence for the Level 1-Level 2 Distinction. *Developmental Psychology*, *17*(1), 99–103.

Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance-reality distinction. *Cognitive Psychology*, *15*(1), 95–120. https://doi.org/10.1016/0010-0285(83)90005-1

Flavell, J. H., Shipstead, S. G., & Croft, K. (1978). Young Children's Knowledge about Visual Perception: Hiding Objects from Others. *Child Development*, *49*(4), 1208–1211.

Fodor, J. A. (1992). A theory of the child's theory of mind. *Cognition*, *44*(3), 283–296. https://doi.org/10.1016/0010-0277(92)90004-2

Friedman, O., Griffin, R., Brownell, H., & Winner, E. (2003). Problems with the Seeing = Knowing Rule. *Developmental Science*, *6*(5), 505–513. https://doi.org/10.1111/1467-7687.00308

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, *10*(4), 483–527. https://doi.org/10.1016/0885-2014(95)90024-1

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*, *134*(1), 31–60. https://doi.org/10.1037/0033-2909.134.1.31

Gautam, S., Suddendorf, T., Henry, J. D., & Redshaw, J. (2019). A taxonomy of mental time travel and counterfactual thought: Insights from cognitive development. *Behavioural Brain Research*, *374*, 112108. https://doi.org/10.1016/j.bbr.2019.112108

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

Grice, P. (1975). *Logic and Conversation.pdf*.

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

Grueneisen, S., Wyman, E., & Tomasello, M. (2015). "I Know You Don't Know I Know..." Children Use Second-Order False-Belief Reasoning for Peer Coordination. *Child Development*, *86*(1), 287–293. https://doi.org/10.1111/cdev.12264

Guajardo, N. R., & Cartwright, K. B. (2016). The contribution of theory of mind, counterfactual reasoning, and executive function to pre-readers' language comprehension and later reading awareness and comprehension in elementary school. *Journal of Experimental Child Psychology*, *144*, 27–45. https://doi.org/10.1016/j.jecp.2015.11.004

Guttentag, R., & Ferrell, J. (2008). Children's understanding of anticipatory regret and disappointment. *Cognition & Emotion*, *22*(5), 815–832. https://doi.org/10.1080/02699930701541542

Halberda, J., Cesana-Arlotti, N., Jauregui, S., & Nichols, S. (2023). *Developmental change in* *children's understanding of possible and impossible events in a "plinko" physical probability device*. Unpublished Manuscript.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(6), 803–864. https://doi.org/10.1017/S0140525X98001769

Happé. (1993). Cognition—Communicative competence & theory of mind in autism—A test of relevance theory—Francesca Happé 48 (1993) 101-119 .pdf. *Cognition*, *48*, 101–119.

Happé, F. G. E. (1994). An Advanced Test of Theory of Mind: Understanding of Story Characters' Thoughts and Feelings by Able Autistic, Mentally Handicapped, and Normal Children and Adults 1. *Journal of Autism and Developmental Disorders*, *24*(2).

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*(3), 233–259. https://doi.org/10.1016/S0010-0277(96)00715-9

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). *The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? 298*.

Hedger, J. A., & Fabricius, W. V. (2011). True Belief Belies False Belief: Recent Findings of Competence in Infants and Limitations in 5-Year-Olds, and Implications for Theory of Mind Development. *Review of Philosophy and Psychology*, *2*(3), 429–447. https://doi.org/10.1007/s13164-011-0069-9

Henzi, S. P., de Sousa Pereira, L. F., Hawker-Bond, D., Stiller, J., Dunbar, R. I. M., & Barrett, L. (2007). Look who's talking: Developmental trends in the size of conversational cliques. *Evolution and Human Behavior*, *28*(1), 66–74. https://doi.org/10.1016/j.evolhumbehav.2006.07.002

Hoerl, C., & McCormack, T. (2019). Temporal updating, temporal reasoning, and the domain of time. *Behavioral and Brain Sciences*, *December*. https://doi.org/10.1017/S0140525X19001195

Hogrefe, G., Wimmer, H., & Perner, J. (1986). Ignorance versus False Belief: A Developmental Lag in Attribution of Epistemic States Author ( s ): G. -Juergen Hogrefe , Heinz Wimmer and Josef Perner Ignorance versus False Belief: A Developmental Lag in Attribution of Epistemic States. *Child Development*, *57*(3), 567–582.

Huemer, M., Perner, J., & Leahy, B. (2018). Mental files theory of mind: When do children consider agents acquainted with different object identities? *Cognition*, *171*, 122–129. https://doi.org/10.1016/j.cognition.2017.10.011

Huemer, M., Schröder, L. M., Leikard, S. J., Gruber, S., Mangstl, A., & Perner, J. (2023). The knowledge (“true belief”) error in 4- to 6-year-old children: When are agents aware of what they have in view? *Cognition*, *230*, 105255. https://doi.org/10.1016/j.cognition.2022.105255

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology*, *16*(2), 233–253. https://doi.org/10.1111/j.2044-835X.1998.tb00921.x

Hughes, C. (2016). Theory of mind grows up: Reflections on new research on theory of mind in middle childhood and adolescence. *Journal of Experimental Child Psychology*, *149*, 1–5. https://doi.org/10.1016/j.jecp.2016.01.017

Iao, L.-S., & Leekam, S. R. (2014). Nonspecificity and theory of mind: New evidence from a nonverbal false-sign task and children with autism spectrum disorders. *Journal of Experimental Child Psychology*, *122*, 1–20. https://doi.org/10.1016/j.jecp.2013.11.017

Immel, A.-S., Altgassen, M., Meyer, M., Endedijk, H. M., & Hunnius, S. (2022). Self-projection in early childhood: No evidence for a common underpinning of episodic memory, episodic future thinking, theory of mind, and spatial navigation. *Journal of Experimental Child Psychology*, *223*, 105481. https://doi.org/10.1016/j.jecp.2022.105481

Kampis, D., Kármán, P., Csibra, G., Southgate, V., & Hernik, M. (2021). A two-lab direct replication attempt of Southgate, Senju and Csibra (2007). *Royal Society Open Science*, *8*(8), 210190. https://doi.org/10.1098/rsos.210190

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory of MInd deficits and causal attributions. *British Journal of Psychology*, *89*, 191–204.

Kloo, D., & Perner, J. (2003). Training Transfer Between Card Sorting and False Belief Understanding: Helping Children Apply Conflicting Descriptions. *Child Development*, *74*(6), 1823–1839. https://doi.org/10.1046/j.1467-8624.2003.00640.x

Kloo, D., & Perner, J. (2005). Disentangling dimensions in the dimensional change card-sorting task. *Developmental Science*, *8*(1), 44–56. https://doi.org/10.1111/j.1467-7687.2005.00392.x

Kratzer, A. (1981). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, *10*, 201–216.

Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit Theory of Mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PLOS ONE*, *14*(3), e0213772. https://doi.org/10.1371/journal.pone.0213772

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

Leahy, B. P., & Carey, S. E. (2020). The Acquisition of Modal Concepts. *Trends in Cognitive Sciences*, *24*(1), 65–78. https://doi.org/10.1016/j.tics.2019.11.004

Leahy, B., & Zalnieriunas, E. (2021). *Might and might not: Children's conceptual development and the acquisition of modal verbs*. 20.

Lecce, S. (2021). Theory of mind goes to school. In R. T. Devine & S. Lecce (Eds.), *Theory of Mind in Middle Childhood and Adolescence* (1st ed., pp. 169–192). Routledge. https://doi.org/10.4324/9780429326899-11

Lecce, S., Bianco, F., Devine, R. T., & Hughes, C. (2017). Relations between theory of mind and executive function in middle childhood: A short-term longitudinal study. *Journal of Experimental Child Psychology*, *163*, 69–86. https://doi.org/10.1016/j.jecp.2017.06.011

Lecce, S., Ronchi, L., Del Sette, P., Bischetti, L., & Bambini, V. (2019). Interpreting physical and mental metaphors: Is Theory of Mind associated with pragmatics in middle childhood? *Journal of Child Language*, *46*(2), 393–407. https://doi.org/10.1017/S030500091800048X

Leekam, S., Perner, J., Healey, L., & Sewell, C. (2008). False signs and the non-specificity of theory of mind: Evidence that preschoolers have general difficulties in understanding representations. *British Journal of Developmental Psychology*, *26*(4), 485–497. https://doi.org/10.1348/026151007X260154

Leekam, S. R., & Perner, J. (1991). Does the autistic child have a metarepresentational deficit? *Cognition*, *40*(3), 203–218. https://doi.org/10.1016/0010-0277(91)90025-Y

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

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, *43*(3), 225–251. https://doi.org/10.1016/0010-0277(92)90013-8

Liddle, B., & Nettle, D. (2006). Higher-Order Theory of Mind and Social Competence in school-age Children. *Journal of Cultural and Evolutionary Psychology*, *4*, 215–229. https://doi.org/10.1556/JCEP.4.2006.3

Lockl, K., & Schneider, W. (2007). Knowledge About the Mind: Links Between Theory of Mind and Later Metamemory. *Child Development*, *78*(1), 148–167. https://doi.org/10.1111/j.1467-8624.2007.00990.x

Marcovitch, S., O'Brien, M., Calkins, S. D., Leerkes, E. M., Weaver, J. M., & Levine, D. W. (2015). A longitudinal assessment of the relation between executive function and theory of mind at 3, 4, and 5 years. *Cognitive Development*, *33*, 40–55. https://doi.org/10.1016/j.cogdev.2014.07.001

Matthews, D. (2014). *Pragmatic Development in First Language Acquisition*. https://doi.org/10.1075/clscc.7

McCormack, T., & Feeney, A. (2015). The development of the experience and anticipation of regret. *Cognition & Emotion*, *29*(2), 266–280. https://doi.org/10.1080/02699931.2014.912201

McCormack, T., Ho, M., Gribben, C., O'Connor, E., & Hoerl, C. (2018). The Development of Counterfactual Reasoning about Doubly-Determined Events. *Cognitive Development*, *45*, 1–9.

McCormack, T., O'Connor, E., Cherry, J., Beck, S. R., & Feeney, A. (2019). Experiencing regret about a choice helps children learn to delay gratification. *Journal of Experimental Child Psychology*, *179*, 162–175. https://doi.org/10.1016/j.jecp.2018.11.005

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

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological Bulletin*, *135*(5), 749–773. https://doi.org/10.1037/a0016854

Mitchell, P. (1996). *Acquiring a conception of mind: A review of psychological research and theory.* Routledge.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, *154*, 40–48. https://doi.org/10.1016/j.cognition.2016.05.012

Moll, H., Meltzoff, A. N., Merzsch, K., & Tomasello, M. (2013). Taking versus confronting visual perspectives in preschool children. *Developmental Psychology*, *49*(4), 646–654. https://doi.org/10.1037/a0028633

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

Moses, L. J., & Tahiroglu, D. (2010). Clarifying the Relation between Executive Function and Children's Theories of Mind. In B. Sokol, U. Muller, J. Carpendale, A. Young, & G. Iarocci (Eds.), *Self- and Social-Regulation* (1st ed., pp. 218–233). Oxford University PressNew York. https://doi.org/10.1093/acprof:oso/9780195327694.003.0009

Norbury, C. F. (2005). The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British Journal of Developmental Psychology*, *23*(3), 383–399. https://doi.org/10.1348/026151005X26732

O'Connor, E., McCormack, T., Beck, S. R., & Feeney, A. (2015). Regret and adaptive decision making in young children. *Journal of Experimental Child Psychology*, *135*, 86–92. https://doi.org/10.1016/j.jecp.2015.03.003

O'Connor, E., McCormack, T., & Feeney, A. (2012). The development of regret. *Journal of Experimental Child Psychology*, *111*(1), 120–127. https://doi.org/10.1016/j.jecp.2011.07.002

Oesch, N., & Dunbar, R. I. M. (2017). The emergence of recursion in human language: Mentalising predicts recursive syntax task performance. *Journal of Neurolinguistics*, *43*, 95–106. https://doi.org/10.1016/j.jneuroling.2016.09.008

O'Grady, C., Kliesch, C., Smith, K., & Scott-Phillips, T. (2015). The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior*, *144*(4), 744–763. http://dx.doi.org/10.1037/xge0000076

Oh, S., & Lewis, C. (2008). Korean Preschoolers' Advanced Inhibitory Control and Its Relation to Other Executive Skills and Mental State Understanding. *Child Development*, *79*(1), 80–99. https://doi.org/10.1111/j.1467-8624.2007.01112.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

Onishi, K. H., & Baillargeon, R. (2005). *Do 15-Month-Old Infants Understand False Beliefs? 308*, 5.

Osterhaus, C., & Bosacki, S. L. (2022). Looking for the lighthouse: A systematic review of advanced theory-of-mind tests beyond preschool. *Developmental Review*, *64*, 101021. https://doi.org/10.1016/j.dr.2022.101021

Osterhaus, C., & Koerber, S. (2021). The Development of Advanced Theory of Mind in Middle Childhood: A Longitudinal Study From Age 5 to 10 Years. *Child Development*, *92*(5), 1872–1888. https://doi.org/10.1111/cdev.13627

Osterhaus, C., & Koerber, S. (2022). The complex associations between scientific reasoning and advanced theory of mind. *Child Development*, cdev.13860. https://doi.org/10.1111/cdev.13860

Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of Advanced Theory-of-Mind Tasks. *Child Development*, *87*(6), 1971–1991. https://doi.org/10.1111/cdev.12566

Osterhaus, C., Kristen-Antonow, S., Kloo, D., & Sodian, B. (2022). Advanced scaling and modeling of children's theory of mind competencies: Longitudinal findings in 4- to 6-year-olds. *International Journal of Behavioral Development*, *46*(3), 251–259.

https://doi.org/10.1177/01650254221077334

Papafragou, A., Cassidy, K., & Gleitman, L. (2007). When we think about thinking: The acquisition of belief verbs. *Cognition*, *105*(1), 125–165. https://doi.org/10.1016/j.cognition.2006.09.008

Parkin, L. J. (1994). *Children's understanding of misrepresentation* [University of Sussex]. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.260822

Peloquin, C. M., McDermott, C. H., & Moses, L. J. (2022). *Executive Abilities: A Meta-Analytic Review of Evidence from School-Age Children*.

Perner, J. (1991). *Understanding the representational mind*. The MIT Press.

Perner, J., Brandl, J., & Garnham, A. (2003). What is a perspective problem Developmental issues in belief ascription and dual identity. *Facta Philosophica*, *5*(2), 355–378.

Perner, J., & Horn, R. (2003). Ignorance or False Negatives: Do Children of 4 to 5 Years Simulate Belief With "Not Knowing = Getting it Wrong?" *Journal of Cognition and Development*, *4*(3), 263–273. https://doi.org/10.1207/S15327647JCD0403_02

Perner, J., Huemer, M., & Leahy, B. (2015). Mental files and belief: 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., Kain, W., & Barchfeld, P. (2002). Executive control and higher-order theory of mind in children at risk of ADHD. *Infant and Child Development*, *11*(2), 141–158. https://doi.org/10.1002/icd.302

Perner, J., Kloo, D., & Gornik, E. (2007). Episodic memory development: Theory of mind is part of re-experiencing experienced events. *Infant and Child Development*, *16*(5), 471–490. https://doi.org/10.1002/icd.517

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences*, *3*(9), 337–344. https://doi.org/10.1016/S1364-6613(99)01362-5

Perner, J., & Leahy, B. (2016). Mental Files in Development: Dual Naming, False Belief, Identity and Intensionality. *Review of Philosophy and Psychology*, *7*(2), 491–508. https://doi.org/10.1007/s13164-015-0235-6

Perner, J., & Leekam, S. (2008). The Curious Incident of the Photo that was Accused of Being False: Issues of Domain Specificity in Development, Autism, and Brain Imaging. *Quarterly Journal of Experimental Psychology*, *61*(1), 76–89. https://doi.org/10.1080/17470210701508756

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137. https://doi.org/10.1111/j.2044-835X.1987.tb01048.x

Perner, J., Mauer, M. C., & Hildenbrand, M. (2011). Identity: Key to Children's Understanding of Belief. *Science*, *333*(6041), 474–477. https://doi.org/10.1126/science.1201216

Perner, J., Priewasser, B., & Roessler, J. (2018). Perner et al. - 2018—The practical other teleology and its development.pdf. *Interdisciplinary Science Reviews*.

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. (1995). Episodic Memory and Autonoetic Conciousness: Developmental Evidence and a Theory of Childhood Amnesia. *Journal of Experimental Child Psychology*, *59*(3), 516–548. https://doi.org/10.1006/jecp.1995.1024

Perner, J., Stummer, S., Sprung, M., & Doherty, M. (2002). *Theory of Mind finds Piagetian perspective: Why alternative naming comes with understanding belief*. *103*, 1–22.

Perner, J., & Wimmer, H. (1985). " John Thinks That Mary Thinks That. . ." Attribution of Second-Order Beliefs by 5- to 1 O-Year-Old Children. *Journal of Experimental Child Psychology*, 437–471.

Peskin, J., Comay, J., Chen, X., & Prusky, C. (2016). Does Theory of Mind in Pre-kindergarten Predict the Ability to Think About a Reader's Mind in Elementary School Compositions? A Longitudinal Study. *Journal of Cognition and Development*, *17*(3), 396–

417. https://doi.org/10.1080/15248372.2015.1049697

Peterson, C. C., & Siegal, M. (1998). Changing focus on the representational mind: Deaf, autistic and normal children's concepts of false photos, false drawings and false beliefs. *British Journal of Developmental Psychology*, *16*(3), 301–320. https://doi.org/10.1111/j.2044-835X.1998.tb00754.x

Peterson, C. C., & Wellman, H. M. (2019). Longitudinal Theory of Mind (ToM) Development From Preschool to Adolescence With and Without ToM Delay. *Child Development*, *90*(6), 1917–1934. https://doi.org/10.1111/cdev.13064

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The Mind Behind the Message: Advancing Theory-of-Mind Scales for Typically Developing Children, and Those With Deafness, Autism, or Asperger Syndrome: **The Mind Behind the Message**. *Child Development*, *83*(2), 469–485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Pexman, P. M., & Glenwright, M. (2007). How do typically developing children grasp the meaning of verbal irony? *Journal of Neurolinguistics*, *20*(2), 178–196. https://doi.org/10.1016/j.jneuroling.2006.06.001

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, *95*(2), 201–236. https://doi.org/10.1016/j.cognition.2004.08.004

Polyanskaya, I., Eigsti, I.-M., Brauner, T., & Blackburn, P. (2022). Second-Order False Beliefs and Linguistic Recursion in Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *52*(9), 3991–4006. https://doi.org/10.1007/s10803-021-05277-1

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

Pouscoulous, N., & Tomasello, M. (2020). Early birds: Metaphor understanding in 3-year-olds. *Journal of Pragmatics*, *156*, 160–167.

https://doi.org/10.1016/j.pragma.2019.05.021

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Priewasser, B., Roessler, J., & Perner, J. (2013). Competition as rational action: Why young children cannot appreciate competitive games. *Journal of Experimental Child Psychology*, *116*(2), 545–559. https://doi.org/10.1016/j.jecp.2012.10.008

Pylyshyn, Z. W. (1978). Cognition and consciousness in nonhuman species: Open Peer Commentary and Authors' Responses. *The Behavioral and Brain Sciences*, *4*, 592–593.

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition*, *117*(2), 230–236. https://doi.org/10.1016/j.cognition.2010.08.003

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). *Counterfactual Reasoning: Developing a Sense of Nearest possible world*. *81*, 376–389.

Rafetseder, E., & Perner, J. (2012). When the alternative would have been better: Counterfactual reasoning and the emergence of regret. *Cognition & Emotion*, *26*(5), 800–819. https://doi.org/10.1080/02699931.2011.619744

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, *114*(3), 389–404. https://doi.org/10.1016/j.jecp.2012.10.010

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. (2017). Theory of Mind. In *Cambrisge Encyclopedia of Child Development* (Vol. 4, pp. 507–514).

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

Rakoczy, H., Bergfeld, D., Schwarz, I., & Fizke, 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

Rakoczy, H., Harder-Kasten, A., & Sturm, L. (2012). The decline of theory of mind in old age is (partly) mediated by developmental changes in domain-general abilities: The decline of theory of mind in old age. *British Journal of Psychology*, *103*(1), 58–72. https://doi.org/10.1111/j.2044-8295.2011.02040.x

Rakoczy, H., & Oktay-Gür, N. (2020). Why Do Young Children Look so Smart and Older Children Look so Dumb on True Belief Control Tasks? An Investigation of Pragmatic Performance Factors. *Journal of Cognition and Development*, *21*(2), 213–239. https://doi.org/10.1080/15248372.2019.1709467

Rakoczy, H., Warneken, F., & Tomasello, M. (2007). “This way!”, “No! That way!”-3-year olds know that two people can have mutually incompatible desires. *Cognitive Development*, *22*(1), 47–68. https://doi.org/10.1016/j.cogdev.2006.08.002

Redshaw, J. (2014). Does metarepresentation make human mental time travel unique? *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(5), 519–531. https://doi.org/10.1002/wcs.1308

Redshaw, J., & Suddendorf, T. (2016). Children’s and Apes’ Preparatory Responses to Two Mutually Exclusive Possibilities. *Current Biology*, *26*(13), 1758–1762. https://doi.org/10.1016/j.cub.2016.04.062

Redshaw, J., & Suddendorf, T. (2020). Temporal Junctures in the Mind. *Trends in Cognitive Sciences*, *24*(1), 52–64. https://doi.org/10.1016/j.tics.2019.10.009

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12–21. https://doi.org/10.1037/0012-1649.33.1.12

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., & Apperly, I. A. (2006). Children's Sensitivity to Their Own Relative Ignorance: Handling of Possibilities Under Epistemic and Physical Uncertainty. *Child Development*, *77*(6), 1642–1655. https://doi.org/10.1111/j.1467-8624.2006.00964.x

Ruffman, T. (1996). Do Children Understand the Mind by Means of Simulation or a Theory? Evidence From Their Understanding of Inference. *Mind & Language*, *11*(4), 388–414. https://doi.org/10.1111/j.1468-0017.1996.tb00053.x

Russell, J. (1992). The theory theory: So good they named it twice? *Cognitive Development*, *7*(4), 485–519. https://doi.org/10.1016/0885-2014(92)80005-Z

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The Development of Executive Functioning and Theory of Mind: A Comparison of Chinese and U.S. Preschoolers. *Psychological Science*, *17*(1), 74–81. https://doi.org/10.1111/j.1467-9280.2005.01667.x

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain*, *128*(5), 1102–1111. https://doi.org/10.1093/brain/awh464

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, *8*(9), 657–661. https://doi.org/10.1038/nrn2213

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: General*, *141*(3), 433–438. https://doi.org/10.1037/a0025458

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

Searle, J. R. (1969). *Speech acts: An essay on the philosophy of language*. MA: Cambridge University Press.

Searle, J. R. (1983). *Intentionality: An eEsay in the Philosophy of Mind*. Cambridge University Press.

Siegal, M. (1999). Language and thought: The fundamental significance of conversational awareness for cognitive development. *Developmental Science*, *2*(1), 1–14. https://doi.org/10.1111/1467-7687.00048

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

Sperber, D., & Wilson, D. (1986). *The Interpretation of Utterances: Semantics, Pragmatics & Rhetoric*. 338.

Stiller, J., & Dunbar, R. I. M. (2007). Perspective-taking and memory capacity predict social network size. *Social Networks*, *29*(1), 93–104. https://doi.org/10.1016/j.socnet.2006.04.001

Suddendorf, T. (1999). The rise of the metamind. In *The descent of mind: Psychological perspectives on hominid evolution* (Vol. 25, pp. 218–260). Oxford University Press. https://linkinghub.elsevier.com/retrieve/pii/S0960982214013670

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel and is it unique to humans? *Behavioral and Brain Sciences*, *30*, 299–351.

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers Can Attribute Second-Order Beliefs. *Developmental Psychology*, *30*(3), 395–402. https://doi.org/10.1037/0012-1649.30.3.395

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months: Attribution of beliefs to a geometric shape. *British Journal of Developmental Psychology*, *30*(1), 30–44. https://doi.org/10.1111/j.2044-835X.2011.02046.x

Surtees, A. D. R., Butterfill, S. A., & Apperly, I. A. (2012). Direct and indirect measures of Level-2 perspective-taking in children and adults: Level-2 perspective-taking. *British Journal of Developmental Psychology*, *30*(1), 75–86. https://doi.org/10.1111/j.2044-835X.2011.02063.x

Szücs, M. (2014). The role of Theory of Mind, age, and reception of grammar in metaphor and irony comprehension of preschool children. *Proceedings of the Third Central European Conference for Postgraduate Students. Budapest: Pázmány Péter Catholic University*, 98–113.

Tulving, E. (1985). How Many Memory Systems Are There? *American Psychologist*.

Tulving, E. (2005). Episodic Memory and Autonoesis: Uniquely Human? In *The missing link in cognition: Origins of self-reflective consciousness* (pp. 3–56). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195161564.003.0001

Uprichard, B., & McCormack, T. (2019). Becoming Kinder: Prosocial Choice and the Development of Interpersonal Regret. *Child Development*, *90*(4), e486–e504. https://doi.org/10.1111/cdev.13029

Valle, A., Massaro, D., Castelli, I., & Marchetti, A. (2015). Theory of Mind Development in Adolescence and Early Adulthood: The Growing Complexity of Recursive Thinking Ability. *Europe's Journal of Psychology*, *11*(1), 112–124. https://doi.org/10.5964/ejop.v11i1.829

Van Duijvenvoorde, A. C. K., Huizenga, H. M., & Jansen, B. R. J. (2014). What is and what could have been: Experiencing regret and relief across childhood. *Cognition and Emotion*, *28*(5), 926–935. https://doi.org/10.1080/02699931.2013.861800

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, *191*, 103997. https://doi.org/10.1016/j.cognition.2019.06.009

Weimer, A. A., Warnell, K. R., Ettekal, I., Cartwright, K. B., Guajardo, N. R., & Liew, J. (2021). Correlates and antecedents of theory of mind development during middle childhood and adolescence: An integrated model. *Developmental Review*, *59*, 100945. https://doi.org/10.1016/j.dr.2020.100945

Weisberg, D. P., & Beck, S. R. (2010). Children's thinking about their own and others' regret and relief. *Journal of Experimental Child Psychology*, *106*(2–3), 184–191. https://doi.org/10.1016/j.jecp.2010.02.005

Wellman, H. M. (2018). Theory of Mind Across the Lifespan? *Zeitschrift Für Psychologie*, *226*(2), 136–138. https://doi.org/10.1027/2151-2604/a000330

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., Fang, F., & Peterson, C. C. (2011). Sequential Progressions in a Theory-of-Mind Scale: Longitudinal Perspectives. *Child Development*, *82*(3), 780–792. https://doi.org/10.1111/j.1467-8624.2011.01583.x

Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. A. C., Chevalier, N., & Espy, K. A. (2011). The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology*, *108*(3), 436–452. https://doi.org/10.1016/j.jecp.2010.08.008

Wilson, R., Hruby, A., Perez-Zapata, D. I., van der Kleij, S., & Apperly, I. (2022). *Is Recursive "Mindreading" Really an Exception to Limitations on Recursive Thinking* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qv2hb

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

Winner, E., & Gardner, H. (1993). Metaphor and irony: Two levels of understanding. *Metaphor and Thought*, *2*, 425–443.

Winner, E., Levy, J., Kaplan, J., & Rosenblatt, E. (1988). Children's Understanding of Nonliteral Language. *Journal of Aesthetic Education*, *22*(1), 51. https://doi.org/10.2307/3332964

Wyman, E., Rakoczy, H., & Tomasello, M. (2013). Non-verbal communication enables children's coordination in a "Stag Hunt" game. *European Journal of Developmental Psychology*, *10*(5), 597–610. https://doi.org/10.1080/17405629.2012.726469

Xiao, Y., Geng, F., Riggins, T., Chen, G., & Redcay, E. (2019). Neural correlates of developing theory of mind competence in early childhood. *NeuroImage*, *184*, 707–716. https://doi.org/10.1016/j.neuroimage.2018.09.079

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition*, *35*(1), 41–68. https://doi.org/10.1016/0010-0277(90)90036-J

Zelazo, P. D. (2004). The development of conscious control in childhood. *Trends in Cognitive Sciences*, *8*(1), 12–17. https://doi.org/10.1016/j.tics.2003.11.001

Zelazo, P. D., & Frye, D. (1998). Cognitive Complexity and Control: II. The Development of Executive Function in Childhood. *Current Directions in Psychological Science*, *7*(4), 121–126.

**Appendix A: Schidelko, Schünemann, Rakoczy, & Proft (2021)**

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task. *Frontiers in Psychology*, 12, 4573.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 13 October 2021
doi: 10.3389/fpsyg.2021.703238

# Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task

*Lydia Paulin Schidelko*†, Britta Schünemann†, Hannes Rakoczy and Marina Proft*

*Department of Developmental Psychology, University of Göttingen, Göttingen, Germany*

Recently, online testing has become an increasingly important instrument in developmental research, in particular since the COVID-19 pandemic made in-lab testing impossible. However, online testing comes with two substantial challenges. First, it is unclear how valid results of online studies really are. Second, implementing online studies can be costly and/or require profound coding skills. This article addresses the validity of an online testing approach that is low-cost and easy to implement: The experimenter shares test materials such as videos or presentations via video chat and interactively moderates the test session. To validate this approach, we compared children's performance on a well-established task, the change-of-location false belief task, in an in-lab and online test setting. In two studies, 3- and 4-year-old received online implementations of the false belief version (Study 1) and the false and true belief version of the task (Study 2). Children's performance in these online studies was compared to data of matching tasks collected in the context of in-lab studies. Results revealed that the typical developmental pattern of performance in these tasks found in in-lab studies could be replicated with the novel online test procedure. These results suggest that the proposed method, which is both low-cost and easy to implement, provides a valid alternative to classical in-person test settings.

**Keywords: online studies, validation study, developmental psychology, psychology methods, Theory of Mind, false belief**

OPEN ACCESS

***Edited by:***
*Lisa Oakes, University of California, Davis, United States*

***Reviewed by:***
*Lindsay Bowman, University of California, Davis, United States*
*Rose M. Scott, University of California, Merced, United States*

***Correspondence:**
*Lydia Paulin Schidelko*
*lydiapaulin.schidelko@uni-goettingen.de*

*†These authors share first authorship*

***Specialty section:***
*This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology*

***Received:*** *30 April 2021*
***Accepted:*** *16 September 2021*
***Published:*** *13 October 2021*

***Citation:***
*Schidelko LP, Schünemann B, Rakoczy H and Proft M (2021) Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task. Front. Psychol. 12:703238. doi: 10.3389/fpsyg.2021.703238*

## INTRODUCTION

Developmental research largely depends on collecting data from children. While varying in methods, set-ups and concrete testing sites, so far, most research has been conducted in an interpersonal, face-to-face setting between an experimenter and a child. Thus, with the beginning of the COVID-19 pandemic, most well-established testing routines were suddenly disrupted and the need for new, safe, and contact-free ways to test children for developmental studies arose.

In the last decade, online testing for psychological research already became more and more prominent for adult studies, with several thousand participants taking part in social science experiments every day on platforms like *Amazon Mechanical Turk* (*MTurk*) and *Prolific* (Bohannon, 2016). More recently, developmental researchers have started to establish first online platforms for children, including *Lookit* (Scott and Schulz, 2017; Scott et al., 2017) and *Discoveries Online* (Rhodes et al., 2020), that both use an unmoderated set-up (where children and families do not interact with the researchers), and *TheChildLab.com* (Sheskin and Keil, 2018) that uses a moderated set-up (where the experimenter calls the families via video chat). However, existing platforms and paradigms are not always available for everyone, because of high costs (e.g., for the experimental testing software), programming requirements (e.g., JavaScript), mandatory software downloads or data protection regulations of the software that do not align with the policies of the research institution. Against this background, when we had to close our lab in March 2020, we decided to establish our own moderated testing paradigm for children. In this article, we want to present this novel set-up and validate it as a suitable, safe and broadly accessible tool for online data collection with children.

In our paradigm, we video call families via the software *BigBlueButton* (BBB) and the experimenter then interacts with the children with the help of animated videos or slides. The combination of BBB and screen-sharing comes with several advantages. Concerning the software, BBB is a free, open source, on-premises software. Additionally, once it is established, it comes with low technical requirements both on side of the experimenter as well as the participant as it runs in all common browsers. Furthermore, the servers for BBB are hosted locally, in our case in our institute. Thus, the use of this software allows researchers to adhere to the highest data protection standards, since only the host can access usage and meta-data. Note, however, that while using BBB offers clear advantages, our general set-up is not limited to BBB but is in principle applicable to almost every video chat software that allows screen sharing.

Having set up a technically suitable paradigm, the most pressing question concerns the data quality that can be obtained by testing children with it. Is our moderated online paradigm really appropriate for (remote) data collection? To answer these questions, we wanted to validate our method. Specifically, we tested whether we can conceptually replicate the effects found in in-lab face-to-face settings in analogous studies implemented in our new online paradigm. Importantly, to avoid population-based effects that could explain potential differences between online and in-lab testing, we drew the samples for both paradigms from the same population: our database of parents who had previously given consent to participate with their children in developmental studies. Both samples were thus comparable concerning (a) socio-demographic variables (age and gender were measured, but the sample is also likely to be comparable concerning other socio-demographic variables, e.g., living environment, as the database only includes families living in and around the same city), (b) familiarization with developmental studies (86% of the children participating in an online study participated in at least one other in-person study in our lab before), and (c) incentive structure (we did not directly compensate parents or children for either paradigm).

For the comparison of the two methods, we used a well-established social-cognitive task: the standard false belief (FB) task (Wimmer and Perner, 1983). The FB task is designed to tap children's ability to attribute subjective mental states to others and is generally seen as the litmus test for having a Theory of Mind (ToM). In its standard version, children see a vignette (acted out with puppets) in which an agent puts an object in one of two boxes and leaves the scene. In her absence the object is transferred to the other box and children are then asked to predict where the agent will look for her object upon her return. Results from countless live studies show that children typically start to master this verbal version of the FB task around the age of four, with younger children falsely predicting that the agent will look for her object where it really is (see Wellman et al., 2001). In addition, we administered the structurally analogous true belief (TB) version of the task. Originally designed to control for extraneous task demands in the FB version, recent studies reported a paradoxical picture: once children master the FB task, they begin to fail the TB task. The TB and FB tasks are thus highly negatively correlated between 3 and 5 such that children first pass the TB and fail the FB task and then show the reverse pattern (see Fabricius et al., 2010; Perner et al., 2015; Oktay-Gür and Rakoczy, 2017). This strange effect in the TB task does not seem to document a conceptual limitation, though. One possibility is that it rather reflects children's sensitivity to task pragmatics that they develop on the basis of their growing Theory of Mind. Several studies reveal that the more advanced in ToM children are, the more pragmatically sensitive they become, and the more they get confused by the triviality of the TB test question given the shared perspective of the experimenter and the child ("Why is the experimenter asking me such as stupid question? I guess there must be a more complex answer than the obvious one"; see Oktay-Gür and Rakoczy, 2017; Rakoczy and Oktay-Gür, 2020). In line with the idea of a high pragmatic component of the TB effect, once the task is modified to become less pragmatically confusing (either by converting it into non-verbal format, or by changing the context so that the question now is less trivial) the effect goes away and children perform competently from age 3 onward without any decline in performance. This is highly relevant for present purposes as it shows that the TB test question in its standard version seem to present a very sensitive measure of children' susceptibility to task pragmatics. The TB task therefore lends itself perfectly as a very stringent test for the comparability of live vs. online testing in even subtle respects of verbal interaction and interpretation.

To validate our online set-up, we thus compare children's performance in the two testing formats (in-lab and online): Do the two paradigms lead to comparable results? This question is not trivial. In fact, the existing literature suggest that there are several indicators that (moderated) online testing might indeed lead to different results. On a general level, there is the video deficit effect (VDE): the phenomenon that children solve the same task later and less accurately when the task is presented in a video than when it is presented by a person (Anderson and Pempek, 2005). The VDE has been found for a variety of tasks such as word-learning (e.g., Thierry and Spence, 2004), object-retrieval (e.g., Troseth and DeLoache, 1998) and imitation (e.g., Klein et al., 2006). Recently, it has also been documented for the FB task: 4- and 5-year-old (who usually pass the task) failed to correctly predict the agent's behavior when the story was presented on a video (Reiß et al., 2014, 2019). Furthermore, there is first data from TheChildLab.com concerning moderated online testing more specifically (Sheskin and Keil, 2018). While in general children tested online provided expected answers on a variety of classical tasks, the FB task seemed to be especially hard: Only 9- to 10-year-olds reliably solved the task, while the 5- to

8-year-olds performed at chance level, opening a gap of around 4–5 years compared to standard in-lab testing results.

For the present validation project, we thus collected data on 3- to 4-year-old children's FB and TB understanding in two online studies and compared it to data we obtained from previous in-lab studies with closely matched protocols. Data from in-lab testing was collected pre-COVID and (partly) reanalyzed for the purpose of the current study (for more details, see **Supplementary Material**). In Study 1, we compared children's performance on the standard FB task between the in-lab and online test setting. In Study 2, we widened the focus and tested whether children's more complex performance patterns in TB and FB tasks would differ between in-lab and online test setting.

## STUDY 1

## Methods

### Participants

The final sample includes 112 monolingual German speaking children aged 36–58 months (mean age = 44.28 months; 56 girls; 64 of them tested in an online test session [mean age = 44.22 months; 31 girls (48%)]; 48 [mean age = 44.35 months; 25 girls (52%)][1] in an in-lab test session). Mean age did not differ between settings [$t(110) = 0.120$, $p = 0.905$]. All children live in and around the same medium sized German university town, that is generally characterized by mixed socio-economic backgrounds[2]. Six additional children were tested but not included in data analyses because of uncooperative behavior (online setting: $n = 3$), technical issues during the test session (online setting: $n = 1$), parental interference during the test session (online setting: $n = 1$) and language issues[3] (in-lab setting: $n = 1$). Children in this and the subsequent study were recruited from a databank of children whose parents had previously given consent to experimental participation.

### Design

All children received two trials of a standard change-of-location FB task. The order and direction of location change (from left to right or vice versa) of the trials were counterbalanced. The tasks were presented either as videos in an online testing format or acted-out in an in-lab setting (for comparable scripts and stimuli and a detailed overview of how the online and in-lab tasks were implemented, see **Supplementary Material**).

### Materials and Procedure

***False belief task***

In the FB task (Wimmer and Perner, 1983), Protagonist A (for example, the boy) placed his object (for example, his ball) in one of two boxes (box 1). In his absence, protagonist B (for example, the girl) moved the ball to the other box (box 2) and the experimenter (E) asked the test question "When the boy returns, where will he look for the ball first?" (Correct answer: box 1) (For additional control questions, see **Supplementary Material**).

### Set-Up

***Moderated online study***

In the online test setting, one female experimenter (E) presented the tasks remotely (on a computer screen, no smartphone) via a video conferencing platform (mainly BigBlueButton, in case of technical issues: *Zoom*). During the test session, the child and E communicated via audio and video streaming. The story lines of the FB task were visually implemented as short video clips (created with the animation software *VyondTM © 2021 GoAnimate*). The child watched the video clips while E told the story lines. At the end of each story line, E asked the control and test questions.

***In-lab study***

In the in-lab setting, children were tested in single sessions by two female experimenters in the laboratory. E1 first acted out the FB task with little figures and then asked the control and test questions.

## Results and Discussion

**Figure 1** shows children's performance on the FB test question as a function of age and test setting. In accordance with the literature, we would expect that children's performance on the standard FB task increases with age. If the test setting has an impact on children's performance in this task, there should be a difference between settings most likely in that the effect of age on children's performance should be different between settings.

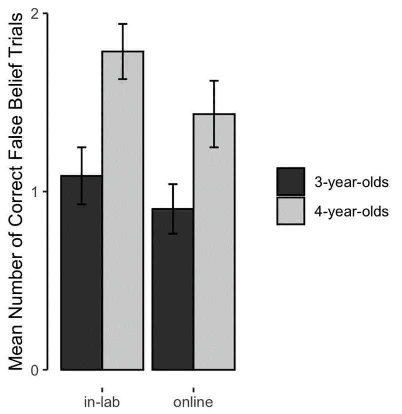



**FIGURE 1** | Mean number of correctly answered false belief trials. This figure depicts the mean number of correctly answered FB trials (0–2) as a function of test setting and age (error bars depict ±1 SE).

[1]The original sample from Schünemann et al. (2022) included sixty-one 2 1/2–4 1/2-year-old children. For the purpose of the current study, we reduced the data set to a relevant subset of children between 3 and 4 1/2 years (for more details, see **Supplementary Material**).

[2]Note that we did not collect any data on race, educational level or socio-economic background.

[3]For the original purpose of the lab study, children were required to be monolingual German (see Schünemann et al., 2022).

For this reason, we set up a Generalized Linear Mixed Model with binomial error structure and a logit link function. As dependent variable, we included children's success on each test trial. To test for an effect of setting and whether the effect of age on performance is different between settings, we included test setting and age measured in months[4] and their interaction. To account for repeated measures, we included children's ID as random intercept effect. We checked for the model's stability by calculating estimates after case wise exclusion of participants. This revealed a stable model. We also checked for multicollinearity (all *VIF*s $\leq 1.001$).

We compared this full model to a null model which included age and the random intercept. This comparison was not significant (likelihood ratio test: $\chi^2 = 0.509$, $df = 2$, $p = 0.775$). Likewise, a closer look at the model revealed that the interaction effect of test setting and age was not significant ($b = -0.840$, $p = 0.567$). Also, the main effect for test setting was not significant ($b = -0.297$, $p = 0.813$). Only the main effect for age was significant ($b = 3.789$, $p = 0.013$). Thus, in accordance with the literature, children's performance increased with age. However, in which setting, in-lab or online, the study was conducted did not impact children's performance.

## STUDY 2

## Methods

### Participants

Seventy-six 36- to 53-month-old native German speaking children were included in the final sample (mean age = 43.76 months; 38 girls). Forty-nine children were tested in an online test setting [mean age = 43.49 months; 23 girls (46%)]. Twenty-seven [mean age = 44.26 months, 15 girls (56%)][5] were tested in an in-lab test setting. Mean age did not differ between settings [$t(74) = 0.605$, $p = 0.547$]. The children live in and around the same medium sized German university town, that is generally characterized by mixed socio-economic backgrounds[6]. Five additional children were tested in the online test setting but excluded from analysis because they were uncooperative ($n = 4$) or had severe language issues (e.g., could not follow the story line and the experimenter's questions; $n = 1$).

### Design

Children again received two trials of a standard change-of-location FB task. Additionally, they received two trials of the TB condition. The two trials of a condition (FB or TB) were presented in blocks. The order of the two blocks and sides of the two trials within the blocks were counterbalanced. The tasks were presented either as an animated slide show in an online testing format or acted-out in an in-lab setting (for comparable scripts and a detailed overview of how the online and in-lab tasks were implemented, see **Supplementary Material**).

[4]Age was z-standardized.

[5]The original sample from Oktay-Gür and Rakoczy (2017, *Exp. 2*) included 171 participants. For the purpose of the current study, we reduced the data set to the relevant subset (for more details, see **Supplementary Material**).

[6]Note that we did not collect any data on race, educational level, or socio-economic background.

### Material and Procedure

#### *False belief and true belief task*

The protocol was slightly adapted from the classic change-of-location task by Wimmer and Perner (1983) used in Study 1 in that E placed the object in the box and moved the object from the first to the second location in the protagonist's absence (FB) or after her return (TB). After that (TB) or after the protagonist's return (FB), E asked the test question "Where does the protagonist think that the toy car is?" [Correct answer: box 1 (FB), box 2 (TB)] (For additional control questions, see **Supplementary Material**).

### Set-Up

#### *Moderated online study*

The same set-up was used as in Study 1. The tasks were presented in a slide show, which was displayed on the child's screen via the platform's screen sharing function. While the child was watching the animated slide show, E told the child the story line and asked the control and test questions.

#### *In-lab study*

In the in-lab format, children were tested as in Study 1 in single sessions by one of five experimenters in the laboratory or in a quiet room of children's day care.

## Results and Discussion

**Figure 2** shows children's performance on the FB (a) and TB (b) test questions as a function of age and test setting. In accordance with the literature, we would expect an interaction between age and the belief type: Children performance on the FB task increases with age while it decreases for the TB task. If the test setting has an impact, this interaction of age and belief type should be different between settings.

Again, we set up a Generalized Linear Mixed Model with binomial error structure and a logit link function and success on test trial as dependent variable. To test for the effect of test setting on the interaction of age and belief type, we included test setting, age and belief type and their interactions in the model. To account for repeated measures, we included children's ID as random intercept effect. The model was stable and not multicollinear (all *VIF*s = 1). Again, we compared this full model to a null model. The null model included age, belief type, their interaction and the random intercept.

This full-null model comparison was not significant (likelihood ratio test: $\chi^2 = 2.312$, $df = 4$, $p = 0.679$). Likewise, a closer look at the model revealed neither a significant 3-way-interaction of test setting, age and belief type ($b = 0.616$, $p = 0.268$), nor any interaction with test setting (with age: $b = -0.196$, $p = 0.606$; with belief type: $b = 0.453$, $p = 0.376$). Also, there was no main effect for test setting ($b = -0.177$, $p = 0.617$). In contrast, the interaction effect of age and belief type was significant ($b = -1.656$, $p < 0.001$). Thus, in accordance with the literature, children's performance increased with age for the FB task and decreased for the TB task. The test setting did not have an impact.

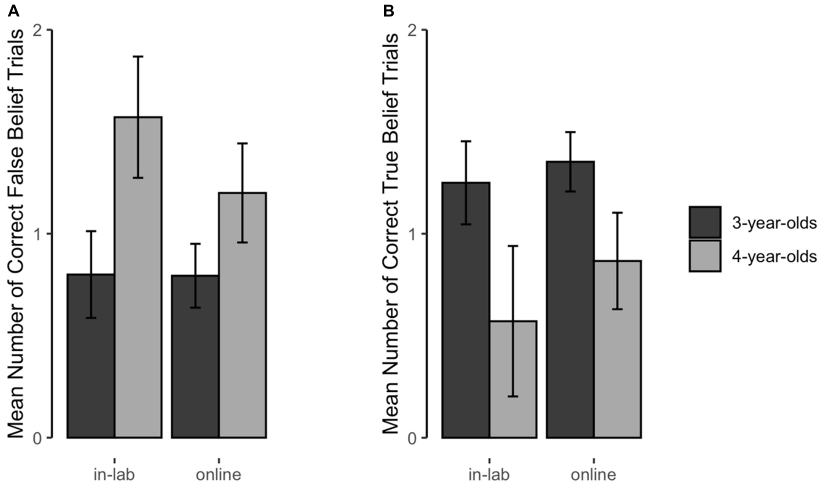



**FIGURE 2 |** Mean number of correctly answered false **(A)** and true **(B)** belief trials. This figure depicts the mean number of correctly answered FB **(A)** and TB **(B)** trials (0–2) as a function of test setting and age (error bars depict ±1 SE).

## GENERAL DISCUSSION

Here, we present and validate a new moderated online testing paradigm for developmental studies. In this paradigm we call families via the video chat software BigBlueButton where the experimenter then interacts with the child with the help of animated videos or slides. The main question regarding the validity of this paradigm was whether it yields results comparable to and converging with in-lab methods. To address this question, we directly compared children's performance in this online paradigm with data from pre-COVID in-lab testing in a standard false belief (FB) and matching true belief (TB) task (Wimmer and Perner, 1983; Oktay-Gür and Rakoczy, 2017). Importantly, we drew samples for both methods from the same database. Thus, all participants were drawn from one population and live in the same local environment. Moreover, in-lab and online samples were matched for age and gender. This reduced potential population-based effects and allowed us to compare the two methods in a very direct and stringent way.

We found no differences between the two testing formats. First, in both studies, 3- and 4-year-old's performance in the online FB task was equivalent to their performance in the acted-out in-lab versions of the task as well as to what we would expect in that age range given the widely documented "4-year-revolution" of mastering standard FB tasks (Perner, 1991; Wellman et al., 2001). Second, in accordance with previous studies, we found a characteristic performance pattern in FB and TB tasks such that children with age become more proficient in the former while becoming less proficient in the latter. This pattern held equally in both testing formats, with no difference between the in-lab and online tests.

By using our moderated online testing paradigm, we thus replicated children's performance from in-lab testing in samples that were drawn from the same population and without facing issues of data loss. Crucially, however, our paradigm does not only seem to closely match interpersonal, face-to-face testing in terms of "cold" indicators such as data quality. Moreover, it also seems to resemble live set-ups in terms of the naturalness and pragmatics of the interaction: when asked a trivial test question (about an agent's true belief), children showed the same response patterns in the online and the live version. One possible interpretation based on recent research (Rakoczy and Oktay-Gür, 2020) is that children were equally prone to draw pragmatic inferences based on their shared perspective with the experimenter, and fall prey to pragmatic confusions in the online setting as in the in-lab setting. In conclusion, our method seems to be a valid and promising instrument for developmental research.

At the same time, the present results leave open many crucial questions. First, in contrast to previous work (e.g., Anderson and Pempek, 2005; Reiß et al., 2014), we found no indication of the video deficit effect (VDE). Thus, watching video presentations (as children did in our Study 1) does not always seem to disrupt children's performance in comparison to live demonstrations. But why didn't the well-documented VDE occur in our paradigm? What is the crucial difference between the cases in which a VDE occurs (in many previous studies) and cases in which it does not (like the present one)? So far, we can only speculate. One crucial difference between the online format and "classical" video presentations is that in our online paradigm the video and the experimenter are both on the screen while in the classic version only the video is presented on-screen with the experimenter sitting next to the child as a live interaction partner who asks test questions. Thus, while in the classic format the child has to handle two parallel worlds (on-screen and live), in the online version all relevant information is presented on-screen, potentially helping the child to encode the video more easily. Other potential influencing factors might be related to the sample (including children's age and their drastically increased familiarity with media use during the pandemic) or the specific task type (see Strouse and Samson, 2021). More future research is needed to systematically test the different conditions under which the VDE occurs in relation to online research.

Second, again in contrast to previous work (Sheskin and Keil, 2018), we found no difference in children's relative performance on belief tasks between online and in-lab settings. Thus, administering the task in a moderated testing paradigm *per se* does not seem to negatively influence children's performance. But then, why were there these gaps in previous work? What is the difference between those cases in which online testing is detrimental to performance and those, like the present one, in which it is not? Again, so far, we can only speculate. When we compare our studies to previous ones, at least two differences emerge: Sheskin and Keil (2018) only presented color coded pictures to the children, whereas we implemented a step-by-step analogous video (or animated slide show) version of the acted-out task version using carefully designed online stimuli [e.g., an animated human hand acting out the change of location and (pre-recorded) verbal interaction between protagonists in the story line onscreen and the experimenter; for more details on scripts and stimuli, see **Supplementary Material**]. This suggests that subtle details of online implementations might matter. Another crucial difference is that we had the opportunity to directly compare the data we obtained from the two methods (online and live) rather than loosely contrasting online data to effects from the literature. For this direct comparison we drew the samples from the same population, while previous studies mostly document a more diverse, broader distributed sample in their online compared to in-lab studies (see also Rhodes et al., 2020). Given these differences, it seems plausible to assume that previous work might have underestimated children's performance in (moderated) online paradigms due to population-based effects. Note, however, that although our samples were drawn from the same population, we cannot exclude selective processes in our studies either. There might be a some sort of selection regarding which parents of our population agreed to online testing. Such processes might have led to a less diverse sample and an overestimation of children's performance. Future research is needed to address this possibility and systematically test for the effects of different population-specific parameters such as socio-economic status, living environment, mobility, closeness to the research institute or time flexibility. For example, even though samples for online and in-lab studies are drawn from the same general population (database), do the sub-samples that respond to live vs. online study invitations differ in subtle demographic respects? Also, note that absence of evidence for a difference between in-lab and online testing of belief tasks, of course, does not amount to evidence of absence of any such potential differences. Future research needs to investigate more systematically and stringently whether there is really no effect of test setting. Such an approach of Bayesian null hypothesis testing will require a larger sample than the current one and will be possible once children can be tested again in an in-lab setting.

The overarching aim of the present project was to find a method for testing children online that is secure, low-cost and easy to implement while yielding comparable results to interpersonal, face-to-face in-lab testing. The results of our two validation studies suggest that with our moderated online testing paradigm we successfully designed such a tool. Future work should now focus on developing the tool further, especially testing its suitability concerning different task types (e.g., more interactive ones that require spontaneous interventions by the child) or different dependent variables (e.g., pointing or eye-tracking). Hopefully, the broader implementation and development of this paradigm then paves the way for more online research in the future, as it has the potential to make developmental research more accessible to a wider audience of participants and researchers.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article are available in the **Supplementary Material**.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the Local Legislation and Institutional Requirements. Written informed consent for participation was not provided by the participants' legal guardians/next of kin because parts of the studies were conducted online. In the online studies, parents/legal guardians gave verbal consent before the testing was started. Verbal consent was recorded and stored separately from the recording of the test session. For the studies conducted in the laboratory or day care, parents/legal guardians gave written consent.

## AUTHOR CONTRIBUTIONS

MP, LS, BS, and HR contributed to conception and design of the study. LS did part of the data collection. BS performed the statistical analysis. MP, LS, and BS wrote the sections and first draft of the manuscript. HR supervised the planning and execution process, provided resources for the data collection, and gave critical review and commentary on the draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 254142454/GRK 2070, Evangelisches Studienwerk Villigst, and Studienstiftung des Deutschen Volkes.

## ACKNOWLEDGMENTS

We thank Jana Rechenburg and Anna Lueb for help with data collection, and Marlen Kaufmann for the organization of data collection.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.703238/full#supplementary-material

## REFERENCES

Anderson, D. R., and Pempek, T. A. (2005). Television and very young children. *Am. Behav. Sci.* 48, 505–522. doi: 10.1177/0002764204271506

Bohannon, J. (2016). Mechanical Turk upends social sciences. *Science* 352, 1263–1264. doi: 10.1126/science.352.6291.1263

Fabricius, W. V., Boyer, T. W., Weimer, A. A., and Carroll, K. (2010). True or false: Do 5-year-olds understand belief? *Dev. Psychol.* 46, 1402–1416. doi: 10.1037/a0017648

Klein, A., Hauf, P., and Aschersleben, G. (2006). The role of action effects in 12-month-olds' action control: A comparison of televised model and live model. *Infant Behav. Dev.* 29, 535–544. doi: 10.1016/j.infbeh.2006.07.001

Oktay-Gür, N., and Rakoczy, H. (2017). Children's difficulty with true belief tasks: competence deficit or performance problem? *Cognition* 162, 28–41. doi: 10.1016/j.cognition.2017.05.002

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Perner, J., Huemer, M., and Leahy, B. (2015). Mental files and belief: A cognitive theory of how children represent belief and its intensionality. *Cognition* 145, 77–88. doi: 10.1016/j.cognition.2015.08.006

Rakoczy, H., and Oktay-Gür, N. (2020). Why Do Young Children Look so Smart and Older Children Look so Dumb on True Belief Control Tasks? An Investigation of Pragmatic Performance Factors. *J. Cogn. Dev.* 1, 1–27. doi: 10.1080/15248372.2019.1709467

Reiß, M., Becker, A., and Krist, H. (2014). Gibt es einen Videodefiziteffekt bei Aufgaben zur Theory of mind? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 46, 155–163. doi: 10.1026/0049-8637/a000112

Reiß, M., Krüger, M., and Krist, H. (2019). Theory of Mind and the Video Deficit Effect: Video Presentation Impairs Children's Encoding and Understanding of False Belief. *Media Psychol.* 22, 23–38. doi: 10.1080/15213269.2017.1412321

Rhodes, M., Rizzo, M. T., Foser-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing Developmental Science via Unmoderated Remote Research with Children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751

Schünemann, B., Schidelko, L. P., Proft, M., and Rakoczy, H. (2022). Children understand subjective (undesirable) desires before they understand subjective (false) beliefs. *J. Exp. Child Psychol.* 213:105268. doi: 10.1016/j.jecp.2021.105268

Scott, K., Chu, J., and Schulz, L. (2017). Lookit (Part 2): Assessing the Viability of Online Developmental Research, Results From Three Case Studies. *Open Mind* 2017:1. doi: 10.1162/OPMI_a_00001

Scott, K. M., and Schulz, L. E. (2017). Lookit: A new online platform for developmental research. *Open Mind* 2017:2. doi: 10.1162/OPMI_a_00002

Sheskin, M., and Keil, F. (2018). TheChildLab.com: a video chat platform for developmental research. *PsyArXiv.* 30:2018. doi: 10.31234/osf.io/rn7w5

Strouse, G. A., and Samson, J. E. (2021). Learning From Video: A Meta-Analysis of the Video Deficit in Children Ages 0 to 6 Years. *Child Dev.* 92, e20–e38. doi: 10.1111/cdev.13429

Thierry, K. L., and Spence, M. J. (2004). A real-life event enhances the accuracy of preschoolers' recall. *Appl. Cogn. Psychol.* 18, 297–309. doi: 10.1002/acp.965

Troseth, G. L., and DeLoache, J. S. (1998). The medium can obscure the message: Young children's understanding of video. *Child Dev.* 69, 950–965. doi: 10.1111/j.1467-8624.1998.tb06153.x

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

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

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

## Appendix B: Schidelko, Huemer, Schröder, Lueb, Perner, & Rakoczy (2022)

Schidelko, L. P., Huemer, M., Schröder, L. M., Lueb, A. S., Perner, J., & Rakoczy, H. (2022). Why Do Children Who Solve FB Tasks Begin to Find True Belief Control Tasks Difficult? A Test of Pragmatic Performance Factors in ToM Tasks. *Frontiers in Psychology*, 12, 797246.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 14 January 2022
doi: 10.3389/fpsyg.2021.797246

Check for updates

# Why Do Children Who Solve False Belief Tasks Begin to Find True Belief Control Tasks Difficult? A Test of Pragmatic Performance Factors in Theory of Mind Tasks

*Lydia P. Schidelko[1*†], Michael Huemer[2,3†], Lara M. Schröder[2,3], Anna S. Lueb[1], Josef Perner[2,3‡] and Hannes Rakoczy[1‡]*

[1] Department of Developmental Psychology, University of Göttingen, Göttingen, Germany, [2] Centre for Cognitive Neuroscience, University of Salzburg, Salzburg, Austria, [3] Department of Psychology, University of Salzburg, Salzburg, Austria

OPEN ACCESS

**Edited by:**
Yuyan Luo,
University of Missouri, United States

**Reviewed by:**
Rose M. Scott,
University of California, Merced,
United States
Andreas Falck,
Lund University, Sweden

***Correspondence:**
Lydia P. Schidelko
lydiapaulin.schidelko@uni-goettingen.de

† These authors share first authorship

‡ These authors share last authorship

**Specialty section:**
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

**Received:** 18 October 2021
**Accepted:** 21 December 2021
**Published:** 14 January 2022

**Citation:**
Schidelko LP, Huemer M, Schröder LM, Lueb AS, Perner J and Rakoczy H (2022) Why Do Children Who Solve False Belief Tasks Begin to Find True Belief Control Tasks Difficult? A Test of Pragmatic Performance Factors in Theory of Mind Tasks.
Front. Psychol. 12:797246.
doi: 10.3389/fpsyg.2021.797246

The litmus test for the development of a metarepresentational Theory of Mind is the false belief (FB) task in which children have to represent how another agent misrepresents the world. Children typically start mastering this task around age four. Recently, however, a puzzling finding has emerged: Once children master the FB task, they begin to fail true belief (TB) control tasks. Pragmatic accounts assume that the TB task is pragmatically confusing because it poses a trivial academic test question about a rational agent's perspective; and we do not normally engage in such discourse about subjective mental perspectives unless there is at least the possibility of error or deviance. The lack of such an obvious possibility in the TB task implicates that there might be some hidden perspective difference and thus makes the task confusing. In the present study, we test the pragmatic account by administering to 3- to 6-year-olds ($N$ = 88) TB and FB tasks and structurally analogous true and false sign (TS/FS) tasks. The belief and sign tasks are matched in terms of representational and metarepresentational complexity; the crucial difference is that TS tasks do not implicate an alternative non-mental perspective and should thus be less pragmatically confusing than TB tasks. The results show parallel and correlated development in FB and FS tasks, replicate the puzzling performance pattern in TB tasks, but show no trace of this in TS tasks. Taken together, these results speak in favor of the pragmatic performance account.

**Keywords: Theory of Mind, pragmatics, true belief, false sign task, knowledge, false belief**

## INTRODUCTION

Theory of Mind (ToM) is the ability to impute mental states, such as beliefs and desires, to oneself and to others (Premack and Woodruff, 1978). The developmental litmus test of ToM are so-called false belief (FB) tasks in which children need to ascribe a mistaken belief to another agent and predict her actions accordingly (Wimmer and Perner, 1983). In the standard change-of-location FB task, the protagonist Maxi puts his chocolate in the blue cupboard and leaves. He does not see that his mother then moves the chocolate to the green cupboard. Children are then asked where Maxi, upon return, will look for his chocolate. Children thus have to represent Maxi's misrepresentation (chocolate in blue cupboard) of the situation (chocolate really in green cupboard). Decades of research with these explicit verbal tasks consistently show that children typically start to ascribe FB around age 4 (Wellman et al., 2001). Success in the FB task goes along with emerging competence in conceptually related tasks that all require metarepresentation – indicating a major conceptual transition in ontogeny (Perner, 1991; Perner and Roessler, 2012).

Before age 4, children systematically fail FB tasks, but pass parallel true belief (TB) control conditions. The TB condition is structurally like the FB condition with the only difference that Maxi watches his mother relocate the chocolate and thus holds a TB[1] about the chocolate's location (Wimmer and Perner, 1983). The test question, like in FB conditions, is variously "Where will Maxi look for his chocolate?" or "Where does Maxi believe his chocolate is?"

The TB condition was devised for younger children who fail the FB condition in order to rule out that FB failure is due to general problems with the narrative structure of the task. Only recently was it administered to a broader age range of children, with puzzling patterns of results. Children who begin to solve the FB task suddenly start to fail the TB task. From age 4 to roughly age 10, then, they systematically answer the TB question incorrectly (predicting that Maxi will erroneously look in the old location; Friedman et al., 2003; Fabricius et al., 2010; Oktay-Gür and Rakoczy, 2017; Rakoczy and Oktay-Gür, 2020; Schidelko et al., 2021).

What do these strange findings mean? One possibility is that they reflect a competence limitation in children's ToM. Contrary to what findings from FB tasks suggest, these findings may be taken as an indication that children do not really engage in metarepresentation until much later (Fabricius et al., 2010, 2021; Hedger and Fabricius, 2011).

Another possibility is that children's difficulty with TB tasks merely reflects pragmatic performance, rather than competence limitations. According to a pragmatic task analysis, TB tasks may be difficult for children from age 4 to 10 because they combine several factors that make the target question pragmatically confusing and thus demanding (Rakoczy and Oktay-Gür, 2020; for related proposals regarding the role of pragmatic factors in FB and other ToM tasks, see, e.g., Siegal and Beattie, 1991; Helming et al., 2014, 2016; Westra, 2016).

*First* of all, the TB question is an academic test question. Regular questions are asked because the speaker herself does not know the answer and requests the missing piece of information from the interlocutor (Searle, 1969). Academic test questions, in contrast, have a much more complex intentional and pragmatic structure: The speaker wants to know whether the interlocutor knows the answer that the speaker knows perfectly well herself. This special question format appears to be difficult to understand for young children (Siegal, 1999).

[1]Commonly considered knowledge that in turn implies a true belief (in contrast to TB tasks that do not imply knowledge; see Gettier, 1963).

*Second*, the TB question is highly trivial: Here, in the story, is a protagonist, who has all the information needed, and now the question is where he will look for an object. The answer is so obvious and common knowledge that it may be difficult to make sense of the corresponding question even if it is understood as an academic test question: Why would someone want to test whether I know what everyone knows?

*Third*, this may be particularly pronounced in the TB case where children are asked where Maxi thinks his chocolate is, or to predict where he will look for it in a situation where he shares common ground and is not subject to any error. Questions in such a context are pragmatically unnatural (Papafragou et al., 2007): We ask for action prediction or explanation or belief ascription only if there is at least the possibility of error and misrepresentation. The test question "What does he believe?" or "What will he do?" therefore suggests that there ought to be an alternative perspective or misrepresentation involved. Yet, the storyline of the TB task does not provide any obvious possibility for error or misrepresentation; children may thus think that they must have missed something and look for a possible alternative perspective on the scenario.

Previous studies have found preliminary evidence for the importance of the first two factors. When tested in a completely non-verbal version of the TB task that removed any (academic and trivial) question, or in a verbal version in which the triviality of the TB question was made explicit ("I'll ask you a baby question"), children between ages 4 and 7 showed no problems with the TB question (Rakoczy and Oktay-Gür, 2020, *Exp. 1 and Exp. 5*).

But what about the third factor that the test question evokes wondering about an alternative perspective? Preliminary evidence comes from one recent study that compared FB/TB tasks with an analogous task that involves non-mental representations, the False Photo (FP) task (based on Zaitchik, 1990). In the FP task, structurally matched to the FB task, an object is put into location 1 and a polaroid camera takes a photo of the scene. While the photo develops, the object is then moved to location 2, and children are asked where the object is in the outdated ("false") photo. Earlier studies revealed that the majority of 3-year-olds failed in both tasks, while the majority of 4-year-olds and older children passed both tasks (with a slightly higher performance in the FB task; Zaitchik, 1990; Leekam and Perner, 1991). Rakoczy and Oktay-Gür (2020, *Exp. 2*) thus used the photo task to explore how asking questions about a rational agent's action or mental perspective may make the TB task pragmatically complex. Four- to 6-year-old children's performance in FB/TB tasks was therefore compared with their performance in analogous False/True Photo tasks. In the new true photo (TP) condition – in close analogy to the TB story – the camera took the photo after the object had already been moved to the new location. Holding the first two factors (trivial academic test question) constant and manipulating only the third (in TB, but not in TP, the test question implicates that there may be an alternative mental representation), the TP condition implements a crucial contrast case: a trivial academic test question about a non-mental representation (of the photo). Consistent with previous findings, 4- to 6-year-old children succeeded in both "false" conditions (FB and FP) whereas their performance in the "true" conditions was markedly different. They showed the previously noted difficulty in the TB tasks (performing below chance), but not in the TP task (performing above chance) (Rakoczy and Oktay-Gür, 2020, *Exp. 2*). These findings thus provide *prima facie* evidence that it really does matter whether trivial academic test questions implicate an alternative perspective due to the agent's mental misrepresentation (rather than referring to non-mental representations).

However, the specific contrast used in that study – between FB and FP tasks – makes the findings difficult to interpret. The reason is that the FP task, strictly speaking, does not involve a misrepresentation. From a theoretical point of view, the "false" photo is actually not false but only outdated (it does not falsely depict the scene at time 2, but depicts the scene as it was at time 1; Perner and Leekam, 2008). Empirically, this analysis is corroborated by findings that FB and FP, though both come to be mastered around the same age, dissociate (fail to correlate) in both typical and atypical development (for an overview, see Perner and Leekam, 2008).

A better task that does involve non-mental misrepresentations is the false sign (FS) task. In this task, in structural analogy to the FB task, a sign post in a story scenario indicates a state of affairs (e.g., that an object is in location 1). The object then moves to a new location (location 2), but the signpost is not changed accordingly and therefore becomes a FS (Parkin, 1994). To solve the task, children need to understand that the actual situation is different from how the sign represents it. Importantly, the sign that shows at time 2 that the object at location 2 is at location 1 is not just outdated (like the photo at time 2 showing that the object was earlier in location 1); it is misleading and false. Empirically, this analysis receives support from a number of studies that suggest that FS and FB tasks are related developmentally in ways in which FB and FP are not: Mastery in both tasks does not only emerge around the same age, they are also highly correlated in typical and atypical development (for an overview, see Perner and Leekam, 2008).

The present study thus capitalizes on this, and develops true and false versions of the sign task (TS/FS) as a minimal contrast to FB/TB tasks in order to test more stringently whether it matters for pragmatics whether trivial academic test questions implicate that there may be alternative mental representations or analogously alternative non-mental representations. The general rationale is the following: If indeed there is a major conceptual transition to metarepresentational thinking around the age of 4, the following pattern of results should be found. Performance in different perspective tasks should show parallel trajectories: younger children tend to fail all tasks requiring an understanding of misrepresentation (e.g., FB and FS), whereas older children tend to master all of them. But if a task poses additional task demands, for example pragmatic factors, no such clear parallel pattern is to be expected. More specifically, if indeed the TB tasks pose pragmatic demands that the TS task lacks (since only in the TB task the test question evokes that there could be an alternative mental representation), we should expect divergent performance: older children worsen in TB but not in TS tasks.

We would thus expect, first, positive correlations of performance between FB and FS tasks and negative correlations between FB and TB tasks as documented in previous studies (e.g., Sabbagh et al., 2006; Leekam et al., 2008 for positive correlations of FB and FS tasks, and Oktay-Gür and Rakoczy, 2017 for negative correlations of FB and TB tasks). Second, we would expect dissociations in performance between TB and TS tasks. Children's performance in the TB task follows the characteristic U-shaped developmental pattern whereas the performance in the TS task will not.

We additionally explored a secondary factor causing wrong answers to the TB question. As in the standard TB task, Maxi watches passively the location change, recent evidence suggests that perhaps children wonder whether Maxi really pays attention and witnesses the location change (Huemer et al., 2019). If children do assume that Maxi did not register the location change, their answer that Maxi will look for the chocolate in its old location would make perfect sense. To reduce any possible ambiguity in this respect Maxi accompanies the location change and we explicitly asked half of the children whether Maxi had seen the location change, as a direct test of whether they have accepted this crucial premise.

## METHODS

### Participants

One-hundred-six 3- to 6-year-old German children participated in the study. They were recruited *via* the platform "KinderSchaffenWissen"[2] and from a databank of children whose parents had previously given consent to experimental participation. The final sample consisted of 88 children (46 female, 42 male; range = 36–83 months, $M$ = 59.3 months), divided into groups of 3-year-olds ($M$ = 42.1 months), 4-year-olds ($M$ = 53.9 months), 5-year-olds ($M$ = 64.5 months), and 6-year-olds ($M$ = 76.7 months), each consisting of 22 children. For more detailed information on participants and exclusion criteria, see **Supplementary Material**.

### Design

Each child received six test trials: two TB and two FB trials (in blocks) and one FS and one TS trial. The order of FB and TB blocks, and of FS and TS trials was counterbalanced. Whether the *Confirmation-of-*Seeing *question* on TB trials was asked or not was varied between participants. The same held true for *Confirmation-of-Change question* on TS trials. For information on the coding procedure, counterbalancing of story plots, number of trials and task protocols, see **Supplementary Material**.

### Procedure

The study was tested in a moderated online setting *via* a video conferencing platform (mainly *BigBlueButton*). During the test session, the child and a female experimenter communicated *via* audio and video streaming. The experimenter presented the tasks as animated stories *via* shared screen.

[2]https://kinderschaffenwissen.eva.mpg.de

### True and False Belief Task

The change-of-location task (after Wimmer and Perner, 1983; used in, e.g., Perner et al., 2011) was presented as an animated slide show in four parallel story lines (see **Figure 1** for task structure). Protagonist 1 placed an object in one of two boxes (B1). Protagonist 2 then transferred the object into the other box (B2) before (TB condition) or after (FB condition) protagonist 1 left the scene. Immediately after protagonist 2 had left the scene, children in the *Confirmation-of-*Seeing *question* condition (TB trials only) were asked:

> *Confirmation-of-*Seeing *question (TB)*: Did [*protagonist 1*] see that? (*Correct answer: yes*)

After that, protagonist 1 returned and children were told that protagonist 1 wanted to have her object now. Then, children were asked the following questions:

> *Test question:* Where will [*protagonist 1*] go now?[3] [*Correct answer: B2 (TB), B1 (FB)*]
> *Memory question:* Where did [*protagonist 1*] put the object in the beginning? (*Correct answer: B1*)
> *Reality question:* Where is the object now? (*Correct answer: B2*)

[3]For the test question in FB/TB we used this German wording: "Sag mal, wo wird Maxi denn jetzt hingehen?," avoiding "look first."

### True and False Sign Task

The sign task (adapted from Parkin, 1994) was presented in two storylines. In a familiarization, children were introduced to the setting and learned that the color of the sign at the crossing indicates the location of the vehicle. In the test trials (see **Figure 1** for task structure), the vehicle drove to location 1 (L1) and the sign showing color 1 was placed at the crossing. After a quick stop at L1 the vehicle drove off again stopped briefly at the crossing, either with (TS condition) or without (FS condition) changing the sign to color 2, and then continued to L2. Half the children were then asked the *Confirmation-of-Change question* on TS trials:

> *Confirmation-of-Change question* (*TS*): Was the sign changed? (Correct answer: yes)

Once the vehicle had stopped at L2, children were asked further questions:

> *Test question*: What does the sign say where the [*vehicle*] is? [*Correct answer: L2 (TS), L1 (FS)*]
> *Reality question*: Where is the [*vehicle*] now? (*Correct answer: L2*)
> *Memory question*: And where was it right before? (*Correct answer: L1*)

Importantly, the sign was a rectangular colored plate (see **Figure 1**) without a directional feature (no arrow). This adaption

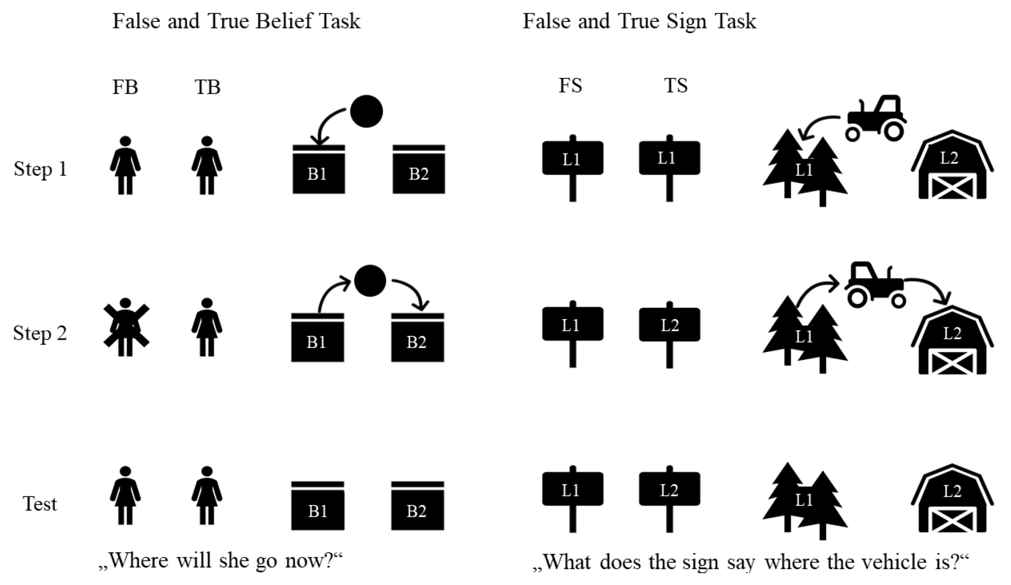



**FIGURE 1 |** Schematic procedure of False/True Belief and False/True Sign Tasks. Children were classified as passers on a given task if they got two out of two correct on belief tasks and one out of one correct on sign tasks (This makes it easier to pass the sign tasks than the belief tasks. We therefore conducted the same analysis as reported below with the first TB and first FB trial only. These analyses show the same results as reported below, see **Supplementary Material**.). They were classified as non-passers otherwise.

was recently introduced for German-speaking populations to ensure that children need to understand the representational feature of the sign (Schuster et al., 2021).

## RESULTS

### Plan of Analysis

The main and novel focus of the present study was on the relation between TB and TS (the former should, while the latter should not show a performance decline after age 4). This focal analysis is only meaningful, however, against the background of two presuppositions: That the patterns of negative correlations between TB and FB and positive correlations between FB and FS performance found in previous studies can be replicated. In preliminary analyses, we therefore tested whether this was fulfilled. We also explored whether posing the *Confirmation-of-*See*ing question* makes a difference to children's TB performance. **Figure 2** provides an overview of children's performance in the various tasks (see **Supplementary Material**, for statistical tests of the developmental trends depicted in **Figure 2** and for an analysis testing the impact of children's gender and the order of presentation).

Data of one six-year-old is missing for the TS condition as, due to an experimental error, they received two FS trails but no TS trial.

### Preliminary Analyses

#### Comparison of True Belief and False Belief Performance

Children showed three patterns of performance: They passed both TB trials and failed at least one FB trial ($M$ = 52.15 months), they passed both FB trials and failed at least one TB trial ($M$ = 66.57 months), they passed both FB and TB trials ($M$ = 65.17 months) (see **Table 1**). Overall, this yields a small to moderate negative correlation between FB and TB tasks (*Pearson's* $r = -0.28$, $p < 0.01$).

Asking the *Confirmation-of-*See*ing question* on TB trials before the test question had no reliable effect, *Chi-squared test* (based on a binomial distribution – children answered either both TB trials correct or not): $p$ = 0.83 (see **Supplementary Material** for further information). Consequently, the two TB conditions (*Confirmation-of-*See*ing question*: yes/no) will be collapsed for all further analyses.

#### Comparison of False Sign and False Belief Performance

False sign and FB task performance showed a moderate to large positive correlation (*Pearson's* $r = 0.46$, $p < 0.001$, see **Table 1**). A *McNemar* test (FB was recoded: children with two correct answers are passers, others are non-passers), revealed no significant difference in the performance of the two tasks ($p$ = 0.69).

### Main Analysis

In contrast to the TB task, performance in the TS was close to ceiling: 93% of the children ($n$ = 82) answered the TS test question correctly (see **Table 2**). A *Chi-squared test* revealed no significant difference in TS performance between children who did and did not receive the *Confirmation-of-Change question* before the test question ($p$ = 0.34). TB and TS tasks were not correlated ($r$ = 0.04, $p$ = 0.68). A *McNemar* test (based on a binomial distribution – children either did or did not consistently pass each task), revealed a significant difference in the performance

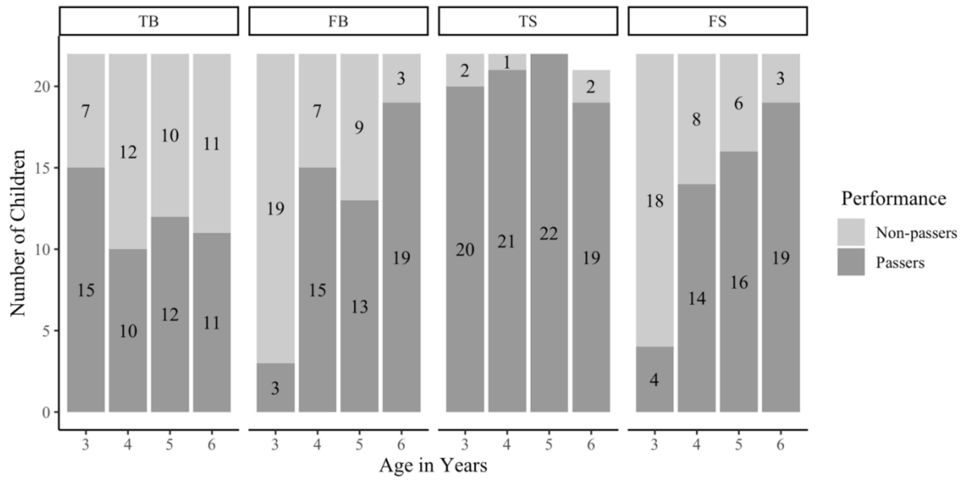



**FIGURE 2 |** Number of children passing and failing true and false versions of the belief and sign task as a function of age.

**TABLE 1 |** Contingency between FB and TB and FS task performance.

| | | TB (correct trials) | | FS (correct trials) | |
|---|---|---|---|---|---|
| | | **0 or 1** | **2** | **0** | **1** |
| FB (correct trials) | 0 or 1 | 13 (7) | 25 (11) | 24 | 14 |
| | 2 | 27 (14) | 23 (12) | 11 | 39 |

*Numbers in parentheses indicate subset of children in the Confirmation-of-Seeing question condition.*

**TABLE 2 |** Contingency between TB and TS task performance.

| | | TS (correct trials) | | |
|---|---|---|---|---|
| | | **0** | **1** | **sum** |
| TB (correct trials) | 0 or 1 | 2 | 38 | 40 |
| | 2 | 3 | 44 | 47 |
| | Sum | 5 | 82 | 87 |

of the two tasks ($p < 0.001$). For further information, see **Supplementary Material**.

## GENERAL DISCUSSION

The aim of the present study was to investigate the source of the puzzling finding that children from around age 4 begin to fail TB tasks. In particular, we tested factors that make the TB task pragmatically demanding and confusing. We compared children's performance in the TB/FB task to performance in the TS/FS task because these tasks are closely matched in structure, involve the same kind of academic and trivial test questions, but contrast in that TB but not TS tasks implicate an alternative mental perspective or misrepresentation.

The main findings were the following: First, the pattern of negative correlations of performance in the TB and FB tasks (Oktay-Gür and Rakoczy, 2017) was replicated. Younger children tended to succeed in the TB but to fail in the FB task whereas older children tended to show the reverse pattern. Second, the convergence and correlation of FB and FS tasks were replicated. Performance in the two tasks that involve a misrepresentation is strongly correlated and develops as found in earlier studies (for an overview, see Perner and Leekam, 2008). This provides additional evidence for a joint developmental transition marking the onset of metarepresentation and perspective understanding (Perner et al., 2002, 2003, 2005; Perner and Roessler, 2012; Moll et al., 2013). Third, there was a marked dissociation between TB and TS tasks. Children showed difficulties in the TB task but not in the TS task. The critical questions in both tasks are trivial academic test questions, but in the former it suggests that there may be an alternative mental representation involved which is not the case in the latter.

Taken together, these results provide new and clear evidence that the TB test question's reference to an agent's rational action (that in turn evokes reference to her subjective perspective) is a crucial part of what makes the question pragmatically odd and thus difficult. In line with previous results[4] (Oktay-Gür and Rakoczy, 2017; Rakoczy and Oktay-Gür, 2020), the present findings corroborate the assumption that it is a combination of various factors that makes TB questions particularly confusing and challenging. They are trivial academic test questions about a rational agent's action. Their triviality and academic nature make them pragmatically odd. Importantly, this effect is particularly strong when the question refers to a rational agent's action. We usually do not ask about a rational agent's perspective or action unless it is unclear what the agent should do or when the agent has a deviant perspective. As this is not obviously the case in the TB question it implicates an alternative hidden mental perspective; trying to figure out what this alternative perspective may be, children venture the guess that the agent might go to the wrong location.

At the same time, the present study leaves open many fundamental questions. First, why exactly is there this sharp difference in the performance of a test question about the belief and action of an agent on the one hand, and the structurally corresponding non-mental representation of an external sign on the other hand? Do we make different kinds of rationality assumptions vis-a-vis the original intentionality of rational agents and the derived intentionality of external signs? In contrast to the intrinsic and original intentionality of agents' mental states, the intentionality of external signs derives from the creators' and users' intentions that confer meaning to them (Searle, 1983). These different kinds of intentionality of mental and non-mental representations might come along with diverging rationality assumptions, and these diverging rationality assumptions may explain the diverging pattern of results in the otherwise structurally analogous tasks. Needless to say that currently this is not more than a speculation; but future studies could and should test for this possibility empirically.

Second, if indeed children find trivial academic test questions about a rational agent's perspective and action confusing, how general is this phenomenon? Here, and in previous research we have shown it for action prediction and (true) belief ascription (Oktay-Gür and Rakoczy, 2017). But would it hold in similar ways, for the ascription of FBs and other types of mental states? For example, is a trivial academic test question about a rational agent's desire pragmatically as confusing? Think of a scenario of the following kind: "Kate has a terrible toothache, but her dentist has the perfect drug for her that is free, has no side effects and immediately makes the pain go away. What will Kate now want to do?" Asking this question might be similarly confusing as asking the test question in the TB task. What is the point of asking about

[4]The present findings from the TB task by themselves are not strictly incompatible with alternative accounts that interpret older children's failure in TB tasks as an indication of lacking ToM competence (Fabricius et al., 2021). According to such accounts, this pattern of results suggests that children do not use metarepresentational belief ascription until much later, but rather operate with a simple (so-called "perceptual access") heuristic. But the results from previous studies (Oktay-Gür and Rakoczy, 2017; Huemer et al., 2019; Study 3; Rakoczy and Oktay-Gür, 2020; Studies 1 and 5) are strictly incompatible with competence limitation accounts. Taken together, these previous findings and the present results, in particular the relations found here between FB and FS tasks, thus provide comprehensive evidence for pragmatic and against competence limitation accounts.

Kate's desire in this situation in which it is completely obvious what is good and to be done? It might make us wonder "Why would someone ask me such obvious things?" and then lead us, in an attempt to make sense of the question, to try out auxiliary assumptions ("Well, perhaps she's a masochist?") It will be an interesting question for future studies to find out whether similar U-shaped curves, based on pragmatic confusion, can be found in mental state ascription more generally.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not provided by the participants' legal guardians/next of kin because the study was conducted online. In the online study, parents/legal guardians gave verbal consent before the testing was started. Verbal consent was recorded and stored separately from the recording of the test session.

## AUTHOR CONTRIBUTIONS

LPS, MH, LMS, JP, and HR contributed to conception and design of the study. AL collected the data. LPS supervised the planning and execution process. HR provided resources for the data collection. LPS performed the statistical analysis. LPS, LMS, and AL wrote the sections and first draft of the manuscript. MH, JP, and HR gave critical review and commentary on the draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

The reported project received financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) by employing LPS on Research Project 254142454/GRK 2070, from the Austrian Science Fund (FWF) by employing MH and LMS on Research Project I 3518-G24 as part of the DACH collaborative project "The structure and development of understanding actions and reasons," and the Austrian Academy of Science (ÖAW) by supporting MH with a DOC scholarship (24691). None of these funding sources had any input in the design of the study, in the collection, analysis, or interpretation of the data, in the writing of the report, or in the decision to submit the article for publication.

## ACKNOWLEDGMENTS

We thank Marlen Kaufmann for the organization of data collection and Hannah Klonk for help with reliability coding.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.797246/full#supplementary-material

## REFERENCES

Fabricius, W. V., Boyer, T. W., Weimer, A. A., and Carroll, K. (2010). True or false: do 5-year-olds understand belief? *Dev. Psychol.* 46:1402. doi: 10.1037/a0017648

Fabricius, W. V., Gonzales, C. R., Pesch, A., Weimer, A. A., Pugliese, J., Carroll, K., et al. (2021). Perceptual access reasoning (PAR) in developing a representational theory of mind. *Monogr. Soc. Res. Child. Dev.* 86, 7–154. doi: 10.1111/mono.12432

Friedman, O., Griffin, R., Brownell, H., and Winner, E. (2003). Problems with the seeing = knowing rule. *Dev. Sci.* 6, 505–513. doi: 10.1111/1467-7687.00308

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis* 23, 121–123. doi: 10.1093/analys/23.6.121

Hedger, J. A., and Fabricius, W. V. (2011). True belief belies false belief: recent findings of competence in infants and limitations in 5-year-olds, and implications for theory of mind development. *Rev. Philos. Psychol.* 2, 429–447. doi: 10.1007/s13164-011-0069-9

Helming, K. A., Strickland, B., and Jacob, P. (2014). Making sense of early false-belief understanding. *Trends Cogn. Sci.* 18, 167–170. doi: 10.1016/j.tics.2014.01.005

Helming, K. A., Strickland, B., and Jacob, P. (2016). Solving the puzzle about early belief-ascription. *Mind Lang.* 31, 438–469. doi: 10.1111/mila.12114

Huemer, M., Gruber, S., Mangstl, A., and Perner, J. (2019). The 'True belief error' in 4- to 6-year-old children. *Int. Conv. Psychol. Sci.*

Leekam, S. R., and Perner, J. (1991). Does the autistic child have a metarepresentational deficit? *Cognition* 40, 203–218. doi: 10.1016/0010-0277(91)90025-Y

Leekam, S., Perner, J., Healey, L., and Sewell, C. (2008). False signs and the non-specificity of theory of mind: evidence that preschoolers have general difficulties in understanding representations. *Br. J. Dev. Psychol.* 26, 485–497. doi: 10.1348/026151007X260154

Moll, H., Meltzoff, A. N., Merzsch, K., and Tomasello, M. (2013). Taking versus confronting visual perspectives in preschool children. *Dev. Psychol.* 49, 646–654. doi: 10.1037/a0028633

Oktay-Gür, N., and Rakoczy, H. (2017). Children's difficulty with true belief tasks: competence deficit or performance problem? *Cognition* 166, 28–41. doi: 10.1016/j.cognition.2017.05.002

Papafragou, A., Cassidy, K., and Gleitman, L. (2007). When we think about thinking: the acquisition of belief verbs. *Cognition* 105, 125–165. doi: 10.1016/j.cognition.2006.09.008

Parkin, L. J. (1994). *Children's Understanding of Misrepresentation*. Doctoral dissertation. Brighton: University of Sussex.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridg, MA: The MIT Press.

Perner, J., Brandl, J. L., and Garnham, A. (2003). What is a perspective problem? Developmental issues in belief ascription and dual identity. *Facta Philos.* 5, 355–378.

Perner, J., and Leekam, S. (2008). The curious incident of the photo that was accused of being false: issues of domain specificity in development, autism, and brain imaging. *Q. J. Exp. Psychol.* 61, 76–89. doi: 10.1080/17470210701508756

Perner, J., Mauer, M. C., and Hildenbrand, M. (2011). Identity: key to children's understanding of belief. *Science* 333, 474–477. doi: 10.1126/science.1201216

Perner, J., and Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends Cogn. Sci.* 16, 519–525. doi: 10.1016/j.tics.2012.08.004

Perner, J., Stummer, S., Sprung, M., and Doherty, M. (2002). Theory of mind finds its Piagetian perspective: why alternative naming comes with understanding belief. *Cogn. Dev.* 17, 1451–1472. doi: 10.1016/S0885-2014(02)00127-2

Perner, J., Zauner, P., and Sprung, M. (2005). "What does "that" have to do with point of view? The case of conflicting desires and "want" in German," in *Why Language Matters for Theory of Mind*, eds J. W. Astington and J. A. Baird (Oxford: Oxford University Press), 220–244.

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526. doi: 10.1017/s0140525x00076512

Rakoczy, H., and Oktay-Gür, N. (2020). Why do young children look so smart and older children look so dumb on true belief control tasks? An investigation of pragmatic performance factors. *J. Cogn. Dev.* 21, 213–239. doi: 10.1080/15248372.2019.1709467

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., and Lee, K. (2006). The development of executive functioning and theory of mind: a comparison of Chinese and US preschoolers. *Psychol. Sci.* 17, 74–81. doi: 10.1111/j.1467-9280.2005.01667.x

Schidelko, L. P., Schünemann, B., Rakoczy, H., and Proft, M. (2021). Online testing yields the same results as lab testing: a validation study with the false belief task. *Front. Psychol.* 12:4573. doi: 10.3389/fpsyg.2021.703238

Schuster, B. A., Huemer, M., Priewasser, B., and Perner, J. (2021). "Preschoolers' understanding of misrepresentation: why the false sign test overestimates performance in german-speaking children," in *Proceedings of the Budapest CEU Conference on Cognitive Development*, Budapest.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Siegal, M. (1999). Language and thought: the fundamental significance of conversational awareness for cognitive development. *Dev. Sci.* 2, 1–14. doi: 10.1111/1467-7687.00048

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

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

Westra, E. (2016). Talking about minds: social experience, pragmatic development, and the false belief task. *Synthese* 194, 4559–4581. doi: 10.1007/s11229-016-1159-0

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

Zaitchik, D. (1990). When representations conflict with reality: the preschooler's problem with false beliefs and "false" photographs. *Cognition* 35, 41–68. doi: 10.1016/0010-0277(90)90036-J

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

## Appendix C: Schidelko, Proft, & Rakoczy (2022)

Schidelko, L. P., Proft, M., & Rakoczy, H. (2022). How do children overcome their pragmatic performance problems in the true belief task? The role of advanced pragmatics and higher-order theory of mind. *PLOS ONE*, 27.

PLOS ONE

RESEARCH ARTICLE

# How do children overcome their pragmatic performance problems in the true belief task? The role of advanced pragmatics and higher-order theory of mind

**Lydia Paulin Schidelko*, Marina Proft, Hannes Rakoczy**

Department of Developmental Psychology, University of Göttingen, Göttingen, Germany

* lydiapaulin.schidelko@uni-goettingen.de



OPEN ACCESS

**Citation:** Schidelko LP, Proft M, Rakoczy H (2022) How do children overcome their pragmatic performance problems in the true belief task? The role of advanced pragmatics and higher-order theory of mind. PLoS ONE 17(4): e0266959. https://doi.org/10.1371/journal.pone.0266959

**Editor:** Jérôme Prado, French National Center for Scientific Research (CNRS) & University of Lyon, FRANCE

**Received:** October 25, 2021

**Accepted:** March 31, 2022

**Published:** April 27, 2022

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0266959



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

## Abstract

The true belief (TB) control condition of the classical location-change task asks children to ascribe a veridical belief to an agent to predict her action (analog to the false belief (FB) condition to test Theory of Mind (ToM) abilities). Studies that administered TB tasks to a broad age range of children yielded surprising findings of a U-shaped performance curve in this seemingly trivial task. Children before age four perform competently in the TB condition. Children who begin to solve the FB condition at age four, however, fail the TB condition and only from around age 10, children succeed again. New evidence suggests that the decline in performance around age four reflects pragmatic confusions caused by the triviality of the task rather than real competence deficits in ToM. Based on these results, it can be hypothesized that the recovery of performance at the end of the U-shaped curve reflects underlying developments in children's growing pragmatic awareness. The aim of the current set of studies, therefore, was to test whether the developmental change at the end of the U-shaped performance curve can be explained by changes in children's pragmatic understanding and by more general underlying developmental changes in recursive ToM or recursive thinking in general. Results from Study 1 ($N$ = 81, 6–10 years) suggest that children's recursive ToM, but not their advanced pragmatic understanding or general recursive thinking abilities predict their TB performance. However, this relationship could not be replicated in Study 2 ($N$ = 87, 6–10 years) and Study 3 ($N$ = 64, 6–10 years) in which neither recursive ToM nor advanced pragmatic understanding or recursive thinking explained children's performance in the TB task. The studies therefore remain inconclusive regarding explanations for the end of the U-shaped performance curve. Future research needs to investigate potential pragmatic and general cognitive foundations of this developmental change more thoroughly.

## Introduction

Theory of Mind (ToM) is the social-cognitive ability to think and reason about one's own and others' mental states [1]. At the core of ToM lies meta-representation: the capacity to represent

**Funding:** The reported project received financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) by employing the first author on Research Project 254142454 / GRK 2070. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

that subjects represent the world in a certain way that can differ from one's own current perspective. The litmus test for ToM is the so-called false belief (FB) task. In this task, participants need to track a story protagonist's belief that comes to differ from reality: Participants see how an object is placed at location 1 in the presence of a protagonist and is then moved from location 1 to a new location 2 in the protagonist's absence. Participants are then asked to ascribe to the protagonist a belief about the object's location ("Where does she think the object is?") or to predict the protagonist's behavior based on her belief ("Where will she look for her object upon her return?) [2]. Individuals with an (explicit) ToM predict that the protagonist will look for her object–based on her false belief–in location 1. Developmentally, children succeed FB tasks around the age of four years. Children younger than four years systematically fail. They predict confidently that the protagonist will look–according to reality–in location 2 [2, 3].

In the history of ToM research, many FB studies also administered an additional True Belief (TB) condition to the children. The TB condition serves as a baseline and control measure to ensure that children, especially younger ones, can cope with the narrative task structure. It is structurally similar to the FB task with the only difference that the protagonist witnesses the object transfer from location 1 to location 2 and thus has a veridical belief about the object's location. Typically, children younger than four years who fail the FB condition, pass the TB condition [3].

Recently, however, the TB condition has been administered to a broader age range of children, with quite puzzling results: with age, children get *worse* in the TB condition. More specifically, children from age four who begin to solve the FB task start to fail the TB task [4–7]. The initial pattern found in younger children–passing the TB condition while failing the FB condition–reverses around this age. Children from age four succeed in the FB condition but fail the TB condition. This performance pattern (younger children pass TB and fail FB and vice versa for older children), reveals itself also at an individual level, in strong negative correlations of the two versions of the tasks. Remarkably, the failure in the TB task persists into late childhood: Only from age seven to ten, performance in the TB task recovers. At this point in development, children pass both FB and TB for the first time. Taken as a whole, the development of performance in the TB task follows a U-shaped trajectory: from high performance in young children around age three to a dramatic decrease around age four when children solve the FB condition to a recovery of performance only in later childhood around age seven to ten. FB performance, in sharp contrast, remains constantly high from around age four [6].

As any U-shaped curve in performance, this unexpected developmental pattern raises at least two fundamental questions: First, how does the decrease in performance in the TB task come about at the beginning of the U-shaped curve? Why do children start to fail TB tasks once they come to master FB tasks? Second, how does the recovery of performance come about at the end of the U-shaped curve? Why do children from around age seven to ten overcome their intermediate difficulty with TB tasks?

One possible answer to both questions is the following: The developmental pattern of the U-shaped performance curve in the TB task does not reflect children's ToM competencies but the development of an understanding of pragmatics. Children's failure in this intermediate state between four and ten is not based on a fundamental problem in ToM understanding but on pragmatic performance limitations [8].

In general, pragmatics pertains to comprehension and production of speech acts and discourse that goes beyond mere literal meaning. For a comprehensive understanding of most speech acts and discourse, additional information besides the literal meanings of the words (sentence meaning) needs to be taken into account in order to determine what is meant (speaker meaning): for example, information about who made an utterance, in which context, against the background of which rules etc.–and in particular, the recipient needs to figure out the speaker's intentions underlying the speech act in question [e.g., 9]. Pragmatics is thus, in some sense, a form of applied ToM.

Regarding the TB performance, it seems quite clear that children from early on do understand the semantics, the literal meaning of the words in the TB test question. Perhaps, however, children in the intermediate state (between ages four and ten) struggle with understanding the use of the test question "Where does the protagonist think the object is?" They do not yet understand what the interlocutor means or wants by asking this question. But why should the TB question be pragmatically challenging? Why do children struggle to grasp the experimenter's intention in asking the TB test question?

A closer inspection of the TB task and the corresponding question reveals that it combines a number of properties that jointly make the task quite peculiar from a pragmatic point of view. First, the TB task is highly trivial: The protagonists clearly sees that the object is moved to a new location and the protagonist, the child and the experimenter share this knowledge about where the object is and everyone knows that the others know, too (this is thus common ground or mutual knowledge). Second, the TB test question is an academic question. The experimenter knows the answer herself. She does not ask this question to gain new information but rather to test whether the child knows the answer, too. Academic question formats are difficult to grasp for young children [10] (for effects of the interviewer's knowledge on children's answer behavior, *see also* [11–13], for related proposals regarding the role of pragmatic factors in FB and other ToM tasks, see, [14–17]. Third, the TB test question asks for a belief ascription or a belief-based action prediction [6]. Normally, we tend to talk about beliefs when we refer to or at least raise the possibility of their falsity [18]. In the TB scenario, however, there is no such obvious possibility. From a purely semantic point of view, the TB test question is utterly unproblematic, indeed highly trivial. Children with a merely literal language use with little sense of pragmatics should thus not find such questions taxing. But for language users with some pragmatic sensitivity, such questions should appear at least *prima facie* odd.

In light of the first question raised above (How does the decrease in performance in the TB task come about at the beginning of the U-shaped curve?), the pragmatic analysis thus yields the following (somewhat simplified) picture: Young children without a sophisticated understanding for ToM are limited in their pragmatic language understanding. They mostly use and interpret language literally (but *see* [19]). Children in this stage of development thus should have no problems with the TB task. However, once children start to develop ToM capacities (i.e., when they pass the FB task), these lay the ground for developing pragmatic understanding [20, 21]. However, their initial ToM and pragmatic understanding are still fragile at this age and their fragile pragmatics then leads them astray in the TB test.

Some evidence speaks in favor of this. First, as reviewed above, performances in the TB and FB tasks are negatively correlated such that children first fail FB and pass TB, and then show the reverse pattern [6]. The performance pattern matches the predictions of the pragmatic analysis such that children's failure in the TB task depends on their success in the FB task: once children develop the prerequisite ToM capacities, they develop an understanding for pragmatics. As a consequence of this development, they suffer from the pragmatic peculiarity of the TB question and fail to answer it correctly while passing the FB task [6]. Second, children succeed in both the TB and FB task after modifications of the task pragmatics in the TB task. For example, children solve the TB task without any decline in the performance curve when the task is presented without or with less trivial language [8]. These results suggest that children show no more difficulties with answering the TB test question correctly once peculiar factors of the tasks are removed. Taken together, first evidence confirms the predictions of the pragmatic analysis at the beginning of the U-shaped performance curve.

The developmental processes at the end of the U-shaped performance curve, however, still raise open questions. How do children overcome their pragmatic difficulties in the TB task later in development, and how does this explain performance recovery at the end of the U-shaped curve? Currently, hardly anything is known about how children overcome their performance limitation in the TB task. From a pragmatic point of view, one possibility is the following: Regarding the beginning of the U-shaped development, the pragmatic analysis predicts that children start to get confused by the peculiarity of the TB test question once they develop sensitivity for pragmatics on the basis of their developing ToM. Applying the same logic to the end of the U-shaped curve, the pragmatic analysis predicts that children succeed again in the TB task once they have undergone further pragmatic development (adults, after all, even though they may find this type of questions funny, have no difficulty in answering it correctly). Taking new steps in pragmatics development might enable children to reason about language use on a higher, more sophisticated level of pragmatic interpretation, and thus to grasp more complex and advanced forms of discourse and speech acts.

Imagine, to illustrate the point, someone asks you, "Did you enjoy the nice weather yesterday?" when it in fact was raining cats and dogs all day. Interpreting this as a regular question asked in order to receive new information would be highly confusing given that the presupposition ("The weather was nice yesterday") is not fulfilled. To understand the actual speaker meaning, you need to stand back from the literal meaning of the question and interpret it at a different, higher level. For example, you might infer the speaker's intention to make an ironic comment about the rainy weather, or you may interpret the question as an academic exercise such that the speaker's intention is to test your knowledge of English past tense. What these examples thus illustrate is that a *prima facie* odd question can make perfect sense when interpreted on a new and higher pragmatic level [22].

Applied to the TB performance, this developmental step might enable children to focus on a new interpretational level that allows them to resolve their initial confusions about the peculiar test question in the TB task ("strange question, but then I'm participating in a study after all; researchers do ask strange questions. . ."). If this idea holds, children who are able to reason at a flexible, higher-order level of pragmatic interpretation should be more likely to answer the TB test question correctly. Accordingly, the emergence of flexible, higher-order pragmatic understanding would determine the end of the U-shaped performance curve.

But what exactly happens in children's pragmatic development at the end of the U-shaped curve? How can the crucial pragmatic development at the end of the U-shaped curve be described, and what are important foundations and correlates of this development?

The pragmatic analysis predicts a developmental progress in advanced pragmatics at the end of the U-shaped curve. If indeed the developmental curve reflects pragmatic progress more generally, this should also become apparent in other areas of advanced pragmatics. To this end, we will compare children's performance in the TB task at the end of the U-shaped curve with developments in *advanced pragmatic understanding* more generally (regarding comprehension of indirect speech acts). The progress in advanced pragmatics, in turn, could itself be rooted in growing *recursive ToM capacities*. And the development in higher-order recursive ToM, in turn, might be based on a more *general ability for recursive thinking*. In the following, we discuss these three possibilities in turn.

## Advanced pragmatic understanding

The crucial step in pragmatic development that leads to the end of the U-shaped curve may reflect a more general phenomenon of developmental progress in pragmatics. Such a general progress in advanced pragmatics might become evident when comparing the performance in the TB task and advanced pragmatics in other areas. Generally, pragmatics is neither a simple and unitary phenomenon [23] nor do pragmatic abilities emerge in simple uniform ways across ontogenetic development [24]. Most relevant for present purposes are *advanced* forms of pragmatic understanding that tend to emerge comparatively late in development. Non-literal language understanding, such as ironic and metaphorical utterances, are prototypical examples [21, 25, 26]. For example, in uttering "It's great weather for our picnic today" when it is raining all day, the speaker does not want the hearer to take her utterance literally. The speaker rather intends the hearer to belief that she thinks that it is not nice to have a picnic outside [27, p.262]. Accordingly, the hearer needs to suppress the initial, literal interpretation to be able to infer the actual meaning, for example, by taking context information into account [24]. Similar processes may be at work in the TB task such that children might overcome their confusion about the trivial test question ("This is too easy, maybe I missed something") by suppressing this initial interpretation and taking context information into account ("I'm participating in a study and the experimenter asks test questions"). Interpreting non-literal language, especially ironic utterances, involves ascriptions of complex intentions and is therefore suggested to be an application of ToM [20, 26]. In neurotypical development, children develop an understanding of metaphors during school age or even preschool age [19, 24, 28]. Its relation to ToM abilities, however, has been controversially discussed in the recent literature [28–30]. Evidence on irony comprehension suggests that children develop an understanding of ironic utterances during school age between six and 13 years of age, depending on the kinds of measures used [24]. The relation of irony understanding to ToM is less controversial: children's performance in irony comprehension correlates with their second-order FB understanding [26, 31].

## Recursive ToM

The crucial foundation of development regarding the end of the U-shaped curve might be even more general than sophisticated pragmatics–for example, the capacity for recursive, higher-order mindreading. The standard ToM tests asks for a first-order mental state ascription, but of course this is not where mental state ascription ends. Advanced forms of ToM enable flexible and higher-order forms of mindreading in which additional levels of mental states can be represented. Ascription of a belief about a belief, for example, constitutes second-order mental state ascription. Virtually, mental states can be recursively embedded within each other infinitely ("A thinks that B thinks that C thinks that D thinks that . . . p"). This understanding for recursively embedded mental states may be the common denominator underlying both advanced pragmatics generally and of the TB performance specifically. It may provide the basis for advanced pragmatics in various forms, for example, by enabling the ascription of higher-order communicative intentions. In the TB task, recursive ToM might enable the ascription of specific higher-order communicative intentions to make sense of the speech act. There are at least two possibilities, how recursive ToM might be involved in the TB task in particular. One possibility is that children who develop an understanding for the pragmatic use of academic test questions (which are also very trivial and pertain to subjective representations) do not any longer get confused by the TB test question at all, but reach the right pragmatic interpretation of the question straight away. They integrate context information and ascribe higher-order intentions and thus make sense of the academic test question from the outset without being confused. A second possibility is that children initially suffer from confusion by the trivial academic TB test question (e.g., "Why does the experimenter ask me such a question? Maybe I must have missed something, maybe the experimenter thinks that I don't know that the protagonists does not know that the object is in that location"). Younger children might not yet be able to resolve this confusion whereas older children, once they have developed recursive mindreading, might be able to ascribe higher-order mental states and thus to resolve the confusion (e.g., "She wants to know whether I know that the protagonist knows that the object is in the new location"). Either way, children's performance on the TB test question should then be related to their recursive ToM capacities more generally.

Recursive ToM reveals itself in various forms, in many different tasks and situations, but tasks that directly test for this understanding for second- or higher-order mental state ascriptions are rare in the literature. Evidence from this line of research shows that children at the age of five to six years can attribute second-order mental states ["A thinks that B believes that . . ."; e.g., 32], but only very little is known about children's development of higher-order ToM beyond second order of recursion. In a study by Liddle and Nettle, for example, ten- to eleven-year-olds performed above chance level in a third-order ToM task and at chance level in a fourth-order ToM task [33]. Adults, in contrast, were able to reason until seventh-order of recursion, in particular, when tested "implicitly" through observing video clips of social interactions compared to explicit measures used in the above reported studies with children [34]. Until now, it therefore remains an open question when exactly children learn to reason about mental states on different levels of embedding and whether this development is fundamental for children's performance in the TB task.

## General ability for recursive thinking

The ability to reason about higher-order mental states, in turn, might be based on a more general ability for recursive thinking. The performance in the TB task, hence, would not only rely on the ability for higher-order ToM but on an even broader ability for recursive operations that is fundamental for higher-order ToM abilities.

Thinking and reasoning recursively is not only of importance in the development of ToM but has been implicated in a number of probably uniquely human abilities such as language, music, mathematics or mental time travel [35–37]. Recursive operations in all these areas require embedding of elements (e.g., mental states, words/clauses, etc.) within elements of the same kind [38]. The corresponding level of reasoning can be more or less clearly defined and quantified (e.g., in the domain of ToM: "A thinks that B thinks that C thinks that p" as third-order mental state ascription).

The general ability for recursive operations might manifest itself in recursive thinking in the various specific areas of application, including higher-order ToM. Consequently, the developmental changes in TB performance might reflect development in advanced pragmatic that builds on recursive ToM that, in turn, is a manifestation of general recursive operations. Once the child has acquired a certain level of general recursive thinking, this enables her–via recursive ToM–to think pragmatically about the TB task at a higher level, overcome her pragmatic confusion and solve the task. If this hypothesized pattern holds, an individual, first, would be able to think recursively on the same level of embedding in different areas of application; and second, the general level of recursive thinking would, at least partly, predict her performance in the TB task.

## Rationale of the present study

In sum, the puzzling developmental pattern of the U-shaped performance curve in the TB task raises two fundamental questions: First, how does the decrease in performance in the TB task come about at the beginning of the U-shaped curve? Second, how does the recovery of performance come about at the end of the U-shaped curve? The pragmatic analysis presents one possible answer to both questions: the U-shaped curve reflects an underlying development in children's understanding of pragmatics.

Previous research has yielded some evidence that speaks to the first question, but so far, no study has empirically addressed the second question. The rationale of the present study, therefore, is to test whether, indeed, the developmental change at the end of the U-shaped performance curve reflects and can be explained by pragmatic development.

To this end, we tested a wide age range of children with the standard FB and TB task. Due to the restrictions of the Covid-19 pandemic, testing was conducted in an online format. We ran the studies as moderated online studies in which the experimenter interacted with the child via video chat while presenting the tasks on screen. With this change in setting, the pragmatic context in which the experimenter administers the TB task to the child was essentially different compared to earlier studies. A preliminary question thus was whether the typical performance curve replicates in the new format in children between six and ten years. The age range is expected to include younger children who still fail and older children who succeed again in the TB task (e.g., [6]), representing the right half of the U-shaped performance curve.

The main research question then was: What are the factors that explain the end of the U-shaped performance curve in the TB task? Here, we explore different possibilities based on the pragmatic analysis introduced above:

1. Is the TB pattern a function of advanced pragmatics development?
2. Even more generally: Is it a function of recursive ToM, or even of recursive thinking in general?

In order to do so, *Advanced Pragmatic Understanding* was operationalized in a task that asked for children's metaphor and irony understanding. *Recursive ToM* was operationalized in tasks that tested children's understanding and production of higher-order mental state ascriptions. The *general ability for recursive thinking* was operationalized as children's recursive language abilities as a proxy for their general recursive abilities. This task tested children's understanding for embedded recursive clauses. Children were tested in these tasks and the TB task in three online studies. Table 1 displays the tasks administered in Study 1–3.

## Study 1

### Method

This research was conducted in accordance with the Declaration of Helsinki and the Ethical Principles of the German Psychological Society (DGPs), the Association of German Professional Psychologists (BDP), and the American Psychological Association (APA). It involved no invasive or otherwise ethically problematic techniques and no deception (and therefore, according to National jurisdiction, did not require a separate vote by a local Institutional Review Board; see the regulations on freedom of research in the German Constitution (§ 5 (3)), and the German University Law (§ 22)). Before the test sessions of Studies 1–3 started, informed consent was obtained from the parents of the subjects.

**Table 1. Tasks included in Studies 1–3 in the order presented in the test session.**

| Task | Study 1 | Study 2 | Study 3 |
|---|---|---|---|
| Recursion in general: recursive language abilities | x | x | x |
| Change-of-location: TB and FB | x | x | x |
| Recursive ToM: production | | x | |
| Recursive ToM: understanding | x | x | x |
| Advanced pragmatics understanding | x | | |

https://doi.org/10.1371/journal.pone.0266959.t001

**Design.** Children in all three studies were tested in a single session (30–45 minutes) by a female experimenter (E). The tasks were presented remotely (on a laptop computer screen or tablet computer screen, no smartphone) in an interactive online study via a video conferencing platform (mainly *BigBlueButton*, in case of connection issues, the test session was shifted to *Zoom*). The tasks were embedded in a video, which was displayed via the conference platform in the middle of the child's screen and required the child to give verbal answers. Next to the video, the child was constantly able to see the webcam video of E and herself, so that the child and E were able to communicate via audio and video streaming during the whole test session (*see* [7] for a validation study of this paradigm). Before the beginning of the test session, the caretaker gave verbal consent to the child's participation in the study and the video and audio recording during the test session. The verbal consent was recorded and stored separately from the recording of the test session. The caretaker and the child were informed that they might abort the participation at any given moment. In the beginning of the test session, E advised the child that she could repeat each question if the child had any comprehension difficulties.

**Participants.** Eighty-one 6- to 10-year-old children (72–131 months, *mean age* = 99.52 months; 41 girls, 40 boys) were included in the final sample. Eight additional children were tested but excluded from data analyses because of technical issues during the test session ($N = 6$), uncooperative behavior ($N = 1$) and concentration deficit resulting in >50% incorrectly answered control question in the change-of-location task ($N = 1$). The age range was chosen so broadly in order to compare children who show and do not show the performance difficulties in the TB task. Participants in this and all subsequent studies were recruited from a database of children whose parents had previously given consent to experimental participation as well as via social media.

**Material.** *Test for syntactic recursion*. The task adapted from Arslan and colleagues (2017) tested for the comprehension of embedded relative clauses in German language [39]. Children saw two rows of animals (upper and lower row) on the screen (Fig 1). Each animal was displayed on a different background color. The children were asked to name the location of a corresponding animal on the screen (e.g., "Where is the cow that strokes a horse?" for the first-order syntactic recursion test question). Children had to refer to the animal's location in naming the corresponding background color and the row in which the animal was placed (e.g., "yellow, upper row"). The test questions containing the relative clauses were scaled from first order until fourth order of syntactic recursion and could be repeated up to four times [39]. For a detailed procedure, *see* S1 File.

*Standard change-of-location task*. The children received four trials of the standard change-of-location tasks with different stimuli [2] implemented in short, animated video clips. Protagonist A and her object O were presented to the child before Protagonist A placed O in one of two boxes (box 1). In her presence (TB condition) or absence (FB condition), protagonist B came into the scene and moved O to the other box (box 2) and the following test and control questions were asked:

- Test question: "Where does Protagonist A think that O is?" [correct answer box 1 for FB, box 2 for TB condition]
- Control Question 1: "In which box was O in the beginning?" [correct answer: box 1]
- Control Question 2: "Where is O now?" [correct answer: box 2]

The TB and FB trials were presented in alternating order beginning with a TB trial and the children saw a frozen still image of the last frame of the scenario (Protagonist A and the two boxes) when answering the questions.

**Fig 1. Material for test question with a first-order syntactic recursion "Where is the cow that strokes a horse?", Correct answer "yellow, upper row".**

https://doi.org/10.1371/journal.pone.0266959.g001

*Recursive ToM task*: *Understanding*. The children heard three stories (partly adapted from [33]) accompanied by animated video clips and were asked to answer test questions about the characters' mental states afterwards. The test questions were scaled from second order to fifth order of mental state recursion and children had to decide which of two sentences was true regarding the story line. The sentences were read out by a voice and displayed with pictures on the screen (Fig 2). To choose one of the two sentences, children could either name the side/ color of the picture on the screen or repeat the sentence.

Example story: the video dilemma (adapted from [33])

This is Sarah and this is Olli. Sarah and Olli are in the same class at school. "Hi, I'm Sarah!', "and I'm Olli". Their teacher is Mrs. Brown. Today Mrs. Brown suggests that Sarah and Olli should bring a video to school tomorrow to watch with the other children. Mrs. Brown also says to them, "Make sure you bring a film that I will like too!" (*Mrs. Brown leaves the scene*). Sarah's favorite videos are pirate videos. Olli's favorite videos are horse films. Which will it be? A pirates or a horse film? Olli says to Sarah, "We just can't decide so I think that we should

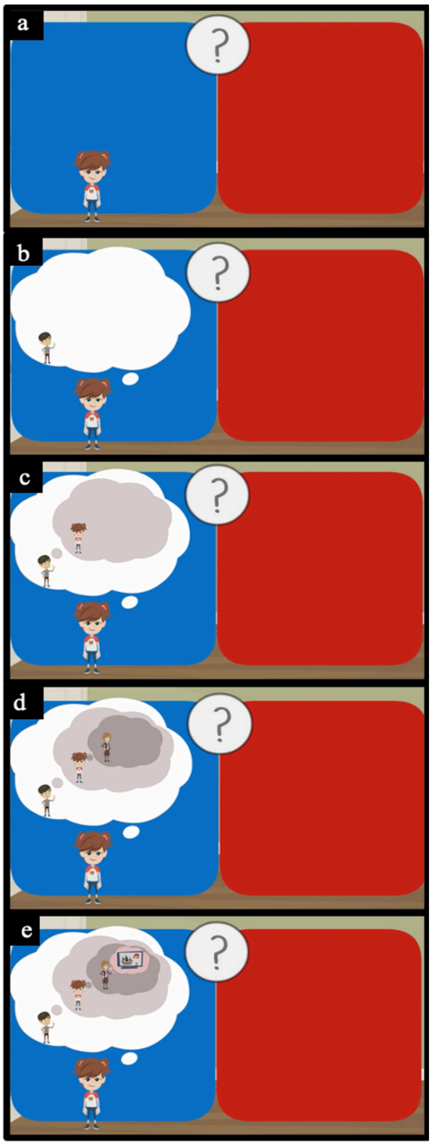



**Fig 2. Screenshots of the visual animation of the fourth-order Theory of Mind test question in the story line "the video dilemma".** *Note*. Children heard a voice slowly reading out the answer sentences while the animation was presented accordingly on the screen. E.g., "Sarah hopes [picture a] that Olli believes [picture b] that she knows [picture c] that Mrs. Brown wants [picture d] them to watch a pirate film [picture e]". After that the second answer option (red side) was read out and presented accordingly.

https://doi.org/10.1371/journal.pone.0266959.g002

take the film that Mrs. Brown would like. Sarah, do you know which one Mrs. Brown would like best?" Sarah is thinking about that. She does not have a clue which film Mrs. Brown would like. But Sarah decides to tell Olli that she knows that Mrs. Brown likes pirate films best. Sarah thinks that this will make Olli agree to take a pirate video to school. Olli listens to this and then Olli says, "We will take a video of pirates then." So, Sarah gets to enjoy her favorite film!

Memory question

Which sentence is true?

a) Sarah likes pirates films best.

b) Sarah likes horse films best.

Test question (ToM Level 4)

Which sentence is true?

a) Sarah hopes that Olli believes that she knows that Mrs. Brown wants that they watch a pirate film.*

b) Sarah hopes that Olli believes that she doesn't know that Mrs. Brown wants them to watch a pirate film.

*German translation with that-complement for want ("möchte, dass")

*Task for advanced pragmatics understanding.* Children received two trials of a pragmatic language task testing for their metaphor and irony understanding [partly adapted from and inspired by 20, 26, 30, 31]. Each trial consisted of a story about two characters accompanied by three pictures (Table 2 and example below).

The questions (and answer options) could be repeated up to four times. During the ironic utterance (Table 2, Picture 3), the speaker's face was not visible to avoid any inferences from their facial expression. To answer the second test question correctly, children had to refer to the speaker's mental state or attitude to the other agent's behavior or refer to the negative outcome of the other agent's behavior or to the opposite/ ironic meaning of the utterance. Answers to this test questions were coded with a fixed coding scheme (adapted from [25, 26], *see* S1 File).

## Results

**Coding of predictors.** In the syntactic recursion task and the recursive ToM understanding task, children were coded with the highest level of recursion until that they performed consistently correct (e.g., child is coded with "3" when she answers the test questions for level 1–3

**Table 2. Example story for Advanced Pragmatics Understanding Task.**

| Picture | Content | Example |
|---|---|---|
| 1 | Information for metaphor | Lisa is running through the apartment all day. She plays with the ball, jumps on the sofa and plays tag with the cat. |
| | Test Question about metaphor | What fits the best? Lisa is. . .<br>• A cloud<br>• A crocodile<br>• A whirlwind (Correct Answer: Metaphor used in German for very active children)<br>• A tree |
| 2 | Story line | In this moment, Lisa is running so fast that she hits the table and all the books fall on the ground. |
| 3 | Ironic utterance | Lisa's older brother enters the rooms and says, "You're very careful today." |
| | Questions about ironic utterance | Test question 1: Does the brother want Lisa to believe that he thinks that she was careful? |
| | | Test question 2: Why does the brother say, "You're very careful today"? |
| | | Control question: Does the brother find that Lisa was careful? |

https://doi.org/10.1371/journal.pone.0266959.t002

correctly, but the test question for level 4 incorrectly). For advanced pragmatics, children received the score of correct trials for the metaphor test questions, the irony test question 1 and irony test question 2 separately (0–2 each). For the coding scheme for irony test question 2 and interrater reliabilities, *see* S1 File.

**Plan of analysis.** In a first step, we assured that the children responded consistently in the two trials of the same condition in the change-of-location task, so that we were able to code children's performance in this task for the subsequent analysis in a binary format (passers vs. non-passers).

Second, in scope of the preliminary analyses, we tested for the typical performance in the TB and FB task in computing comparisons against chance level performance for both TB and FB in the three age groups (young, middle, old). Children of all groups were expected to perform above chance level in the FB task whereas only the oldest age group (9;4–10;11 years) was expected to perform better than chance level performance in the TB task. Additionally, we computed correlations between FB and TB performance which were expected to be negative for the two younger age groups and positive for the oldest age group only.

To address the main research question of factors that influence the performance in the TB task, we computed a logistic regression model. In the logistic regression model, TB performance (passing vs. no-passing) was predicted by recursive syntactic abilities, recursive ToM understanding, advanced pragmatics understanding and children's age. We compared this full model with a control model containing only children's age in months.

**TB and FB performance: Consistency across trials in the standard change-of-location task.** The consistencies in performance of children over the two trials of the same condition of the standard change-of-location task were high. The percentage of children who had two available trials (meaning all control questions answered correctly) and showed the same performance in both trials was 85.90% ($\Phi$ = .68) for the TB trials and 98.75% ($\Phi$ = n/c, due to at least one constant variable) for the FB trials. Therefore, both trials were included in the analysis. For the following analysis, the TB and FB performance were coded as binary variables. Children had to pass both trials of a condition to be assigned to the group of passers. Children failing in one or both trials of a condition were assigned to the group of non-passers.

**Preliminary analyses: TB and FB performances in different age groups of children.** The performance in the change-of-location task as a function of belief type and age is depicted in Fig 3.

To test for the failure in the TB condition and the success in the FB condition of younger children and the success in both conditions in older children, we computed Wilcoxon signed rank tests against chance level performance (0.5) for the three age groups and the two belief conditions. The Wilcoxon tests showed that the youngest age group (6;0–7;7-year-olds) performed significantly above chance in the FB condition ($M$ = .93, p < .0001, r = -.85). The tests could not be computed for the two older age groups due to ceiling effects in the FB condition (7;8–9;3-year-olds and 9;4–10;11-year-olds $M$ = 1). In contrast, Wilcoxon signed rank tests revealed that in the TB condition, only the oldest age group of children (9;4–10;11-year-olds) performed significantly above chance ($M$ = .74, $p$ < .02, $r$ = —.48). Younger children (6;0–7;7-year-olds and 7;8–9;3-year-olds) performed at chance level (6;0–7;7-year-olds: $M$ = .41, $p$ = .34, $r$ = —.18; 7;8–9;3-year-olds: $M$ = .63, $p$ = .18, $r$ = -.26).

The correlation between the TB and FB performance in the change-of-location task is (not-significantly) negative for the whole sample (r(phi) = -.13, $p$ = .24) as well as for the youngest age group (6;0–7;7-year-olds: $r$(phi) = -.34, $p$ = .08). Because of the ceiling effects in the FB condition, the correlation is not computable for the two older age groups.

**Main analyses: Predictors for TB performance.** We removed children failing the first-order FB condition ($N$ = 2) from the following analyses. This was based on the assumption

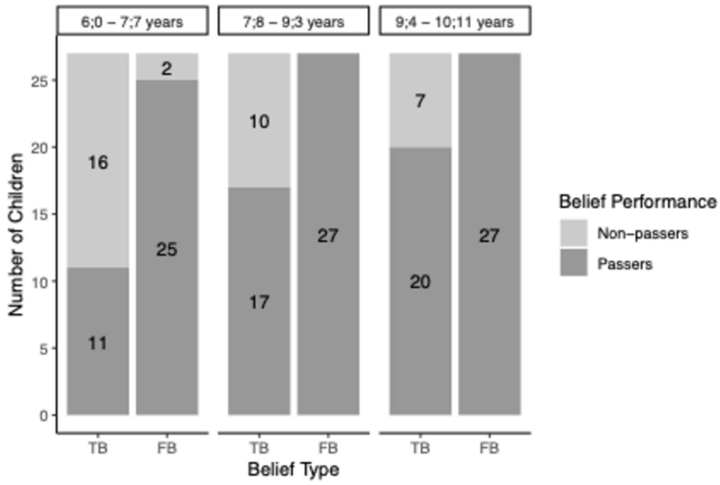



**Fig 3. Children's performances in the standard change-of-location task as a function of age group and belief type in Study 1.**

https://doi.org/10.1371/journal.pone.0266959.g003

that children who still do not succeed in the first-order FB task use different cognitive strategies to solve the TB task compared to the group of children we aim to examine here.

*Descriptive statistics*. The mean performances in the recursive ToM understanding task, the advanced pragmatic language task and the syntactic recursion task as a function of TB performance and age are summarized in Table 3.

For a more detailed summary of answers to irony test question 2, *see* S1 File.

*Logistic regression models*. We estimated the effect of the different predictors of mental state ascription on the TB performance using a multiple logistic regression model. To control for children's age in months, we included it into the model, too. Prior to fitting the model, we checked for the assumptions. We checked for multicollinearity (all *VIF*s $\leq$ 1.38) and linearity of the logit for age ($b$ = 0.08, $p$ = .81), recursive ToM understanding ($b$ = 1.48, $p$ = .73) and syntactic recursion ($b$ = -2.98, $p$ = .27).

We compared the fit of the full model with that of a null model with the control variable only (TB ~ age). As the model comparison is significant, the predictors of mental state

**Table 3. Mean performance (M) and standard deviations (SD) in Syntactic Recursion (Synt. Recurs.), Recursive ToM Understanding (RToM U), metaphor understanding (Metaphor), irony understanding in the first and second test question (Irony1 and Irony2) and for TB non-passers (noTB) and TB passers (TB) in three groups of age.**

| | n | | Synt. Recurs. | | RToM U | | Metaphor | | Irony1 | | Irony2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | M(SD) | | M(SD) | | M(SD) | | M(SD) | | M(SD) | |
| Age | noTB | TB | noTB | TB | noTB | TB | noTB | TB | noTB | TB | noTB | TB |
| 6;0–7;7 | 16 | 11 | 2.19 (1.11) | 2.00 (1.5) | 2.13 (1.36) | 2.44 (0.88) | 1.33 (0.62) | 1.78 (0.67) | 1.44 (0.81) | 1.00 (1.00) | 1.40 (0.83) | 0.89 (0.93) |
| 7;8–9;3 | 10 | 17 | 2.80 (0.92) | 2.94 (1.25) | 2.00 (0.67) | 4.18 (1.19) | 1.80 (0.42) | 1.82 (0.39) | 1.00 (0.94) | 1.35 (0.79) | 1.40 (0.84) | 1.24 (0.83) |
| 9;4–10;11 | 7 | 20 | 2.43 (1.62) | 3.00 (0.97) | 2.57 (1.40) | 3.90 (1.37) | 1.71 (0.49) | 1.80 (0.41) | 1.71 (0.49) | 1.55 (0.60) | 1.86 (0.38) | 1.75 (0.44) |

*Note*. Possible range of performances for recursive ToM understanding: 0–5, for pragmatic language task: 0–2, for syntactic recursion task: 0–4.

https://doi.org/10.1371/journal.pone.0266959.t003

**Table 4. Results of the logistic regression model predicting children's TB performance with their age in months and their performance in tasks of syntactic recursion (Synt. Recurs.), Recursive ToM understanding (RToM U), and Advanced Pragmatics (Metaphor and Irony1 and Irony2 for irony test questions 1 and 2).**

| | *B*(*SE*) | *z* | *p* | 95% CI for Odds Ratio | | |
|---|---|---|---|---|---|---|
| | | | | Lower | Odds Ratio | Upper |
| Included | | | | | | |
| Constant | -4.10 (1.89) | -2.17 | .03* | | | |
| Age in months | 0.02 (0.02) | 1.29 | .20 | 0.99 | 1.02 | 1.06 |
| Synt. Recurs. | -0.19 (0.29) | -0.66 | .51 | 0.46 | 0.83 | 1.44 |
| RToM U | 0.84 (0.23) | 3.67 | < .001*** | 1.54 | 2.33 | 3.84 |
| Metaphor | 0.71 (0.60) | 1.18 | .24 | 0.64 | 2.03 | 6.95 |
| Irony1 | 0.13 (0.39) | 0.34 | .73 | 0.52 | 1.14 | 2.48 |
| Irony2 | -0.68 (0.43) | -1.59 | .11 | 0.21 | 0.51 | 1.15 |

*Note*. $R^2$ = .44 (Nagelkerke). Model $X^2(6) = 24.58$, $p < .001$.

$^{***}p < .001$.

https://doi.org/10.1371/journal.pone.0266959.t004

ascription have an impact on the TB performance (Model $X^2(6) = 24.58$, $p < .001$). More specifically, an increased ability in recursive ToM understanding lead to increased TB performance (B = 0.84, $p < .001^{***}$, OR = 2.33). Pragmatic language abilities, age and syntactic recursion abilities did not affect the TB performance significantly (Table 4).

Fig 4 pictures this difference in performance in recursive ToM understanding between TB-passers and TB-non-passers.

*Post-hoc analyses*. We computed post-hoc one-sided Wilcoxon rank sum tests for TB-passers versus TB-non-passers in recursive ToM understanding as it significantly predicted the outcome in the logistic regression model. Due to multiple testing, Bonferroni correction was applied and resulted in an alpha value of 0.0125 (0.05/4) for this post-hoc computation. The comparison of the performance shows a significant difference in the recursive ToM

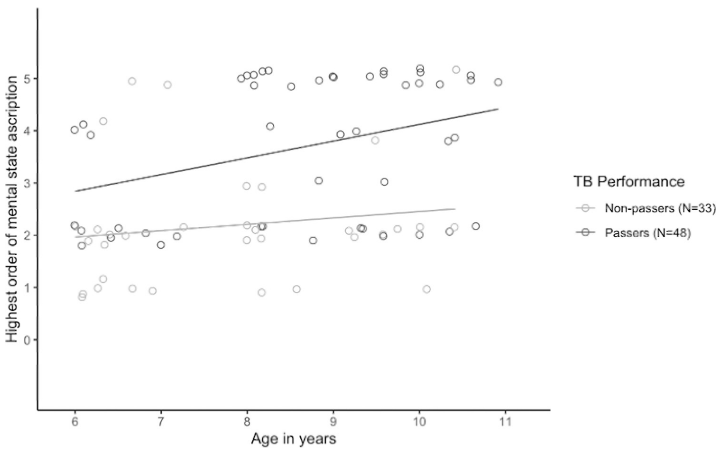



**Fig 4. Children's performance in recursive ToM understanding as a function of their TB performance (passers vs. non-passers) across age.**

https://doi.org/10.1371/journal.pone.0266959.g004

understanding for TB-non-passers (*M* = 1.91, *n* = 33) against TB-passers (*M* = 3.72, *n* = 46, *W* = 319.5, $p < .0001$, *r* = -.53) for the whole sample. For the comparison within the age groups, Wilcoxon rank sum tests reveal that the performance differs significantly for the middle age group between TB-non-passers (*M* = 1.80, *n* = 10) and TB-passers (*M* = 4.18, *n* = 17, *W = 16*, $p < .0001$, *r* = -.73). This difference is not significant for the youngest age group (*M* (noTB) = 1.75, *n* = 16; *M*(TB) = 2.44, *n* = 9, *W* = 50.5, *p* = .09, *r* = -.34) and the oldest age group of children (*M*(noTB) = 2.43, *n* = 7; *M*(TB) = 3.90, *n* = 20, *W* = 34.5, *p* = .02, *r* = -.45).

The achieved power was computed post-hoc for the logistic regression model. For the significant predictor recursive ToM understanding, this resulted in a power of 1 –*ß* = .89.

## Discussion

The expected pattern of typical TB and FB performance in children between six and ten years was replicated in the online study: children in the youngest and middle age group (6;0–9;3 years) failed to perform above chance level in the TB task while FB performance was at ceiling. Only children in the oldest age group (9;4–10;11years) performed proficiently in both conditions. Additionally, the study shows first evidence that children's TB performance can be (partly) explained by their understanding for recursive ToM. However, none of the variables of Advanced Pragmatics understanding or syntactic recursion were significant predictors for children's TB performance.

# Study 2

Study 2, therefore, aimed to replicate this relation between TB performance and children's recursive ToM. In order to explore the underlying recursive ToM abilities in more detail, Study 2 operationalized children's recursive ToM twofold: similar to Study 1, children's understanding for recursive mental state ascriptions was measured. Additionally, children's recursive ToM production was measured to identify the cognitive mechanisms relevant for the TB task more fine-grained.

## Method

**Participants.** The final sample included eighty-seven 6- to 10-year-old children (72–131 months, *mean age* = 101.41 months; 44 girls, 43 boys). Seven additional children were tested but excluded from data analyses because of technical issues during the test session (*N* = 3), experiential error (*N* = 1), uncooperative behavior (*N* = 1), parental interference (*N* = 1) or children's age (child turned out to be too old on the day of the test session; *N* = 1).

**Material.** *Test for syntactic recursion*. The task for syntactic recursion was administered as in Study 1.

*Standard change-of-location task*. For the standard change-of-location task, we used the same material as in Study 1. As the consistency of the two trials (two FB and two TB trials) in Study 1 was high, only one trial per condition was administered in Study 2. Again, the TB trial was always presented first.

*Recursive ToM production*. Children saw two story lines containing bluffs [partly adapted from 40]. In the first story, a character bluffed to mislead someone. In the second story line, the character first uttered a double bluff, i.e. he told the truth as the opponent expected him to lie. As the second story line continued, the same character again uttered a bluff based on his double bluff: he lied as the opponent now expected him to say the truth (triple bluff). The story lines were presented as short animated video clips. After each utterance, children were asked to explain the bluffs ("Why does he say that?"). To receive detailed answers, E asked the child to explain her answer in more detail. The exact wording of this follow-up question depended on the length of the child's initial answer. Answers to the open question were coded binary (child did explain the bluff correctly (1) or did not understand the bluff (0)). Children received the highest score of bluffs that were all explained correctly (0–3).

Example of task for recursive ToM production

Double bluff [shortened, adapted from 40]

The treasure diggers want to find the pirates' treasure. They know that the treasure is either in the field or in the mountains. They hold one of the pirates captive and ask him where the treasure is. The captured pirate is very brave and very smart, he will not let the treasure diggers find the treasure. The treasure is in the mountains. When the treasure diggers ask the pirate where the treasure is, the pirate answers, "in the mountains".

Control question: Is that right what he said? (correct answer: yes)

Test Question: Why did he say that?

For follow-up test question depending on length of the answer to initial test question, *see* S1 File.

*Recursive Tom*: *Understanding*. The task testing for the understanding of recursive ToM was taken over from Study 1 and extended: Children received one additional story and four additional test questions resulting in four story lines, two test questions per story and order of recursion (for more details, *see* S1 File). If a child answered both second order questions or more than 50% of the first four test questions incorrect, the test session was terminated after the third story line.

## Results

**Coding of predictors.** In the syntactic recursion task and the recursive ToM understanding and production task, children were coded with the highest level of recursion up to which they answered the test question/ both test question correctly (e.g., child is coded with "3" for recursive ToM understanding when she answers all test questions for level 1–3 correctly, one test question for level 4 correctly and one test question for level 4 incorrectly). For recursive ToM production, children receive the score of the respective bluff they explained correctly (0–3; "1" when they only explained the bluff correctly, "2" when they explained the bluff and the double-bluff correctly and "3" when they explained all three bluffs correctly). For interrater reliabilities, *see* S1 File.

**Plan of analysis.** The analysis was conducted in close similarity to Study 1. In scope of the preliminary analyses, we tested for the typical performance in the TB and FB task in computing comparisons against chance level performance for both TB and FB in the same three age groups (young, middle, old) and computed correlations between FB and TB performance.

To address the main research question of factors that influence the performance in the TB task, we again computed a logistic regression model. In the logistic regression model, TB performance (passing vs. no-passing) was predicted by children's age, recursive syntactic abilities, recursive ToM production and recursive ToM understanding. We compared this full model with a control model containing only children's age.

**Preliminary analyses: TB and FB performances in different age groups of children.** The performance in the change-of-location task as a function of belief type and age is depicted in Fig 5.

To test for the failure in the TB condition and the success in the FB condition of younger children and the success in both conditions in older children, we computed Wilcoxon signed rank tests against chance level performance (0.5) for the three age groups and the two belief conditions. Wilcoxon showed that the youngest age group (6;0–7;7-year-olds) performed significantly above chance in the FB condition ($M$ = .86, $p < .001$, $r$ = -.70). Wilcoxon tests could

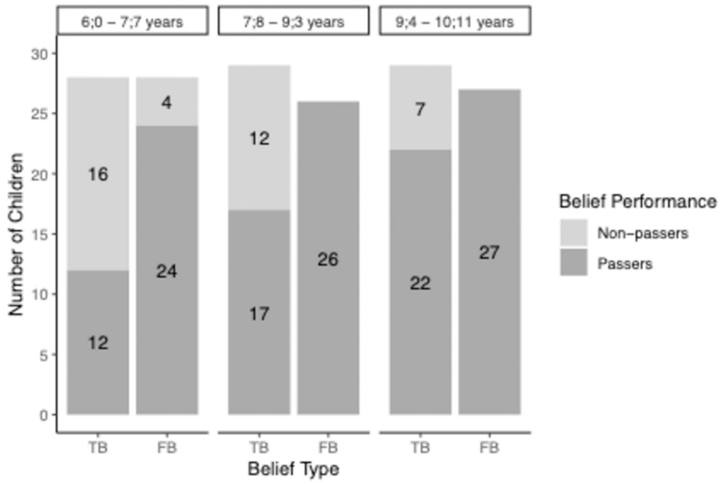



**Fig 5. Children's performance in the change-of-location task as a function of belief type and age in Study 2.** *Note*. Sample size of the groups vary as only children were included who answered the respective control questions correctly.

https://doi.org/10.1371/journal.pone.0266959.g005

not be computed for the two older age groups due to ceiling effects in the FB condition (7;8–9;3-years-olds and 9;4–10;11-year-olds: $M = 1$). In contrast, Wilcoxon signed rank tests revealed that in the TB condition, only the oldest age group of children (9;4–10;11-year-olds) performed significantly above chance ($M = .76$, $p < .01$, $r = -.51$). Younger children (6;0–7;7-year-olds and 7;8–9;3-year-olds) performed at chance level (6;0–7;7-year-olds: $M = .43$, $p = .46$, r = -.14, 7;8–9;3-year-olds: $M = .59$, $p = .36$, $r = -0.17$).

The correlation between the TB and FB performance in the change-of-location task is (not-significantly) negative for the whole sample ($r$(phi) = -.19, $p = .10$) and significantly negative for the youngest age group (6;0–7;7-year-olds: $r$(phi) = -.47, $p = .02$). Because of the ceiling effects in the FB condition, the correlation is not computable for the two older age groups.

**Main analyses: Predictors for TB performance.** *Descriptive statistics*. The mean performances in the recursive ToM understanding task, the recursive ToM production task and the syntactic recursion task as a function of TB performance (passers vs. non-passers) and age (young, middle, old age group) are summarized in Table 5.

**Table 5. Mean performance (M) and standard deviations (SD) in Recursive ToM Understanding (RToM U), Recursive ToM Production (RToM P) and understanding of syntactic recursion (Syntact Recurs.) for TB non-passers and TB passers in three groups of age.**

| | ***n*** | | **Synt. Recurs.** | | **RToM P** | | **RToM U** | |
|---|---|---|---|---|---|---|---|---|
| | | | *M*(*SD*) | | *M*(*SD*) | | *M*(*SD*) | |
| **Age** | **noTB** | **TB** | **noTB** | **TB** | **noTB** | **TB** | **noTB** | **TB** |
| 6;0–7;7 | 16 | 12 | 2.19 (1.11) | 2.08 (1.16) | 1.06 (0.25) | 1.17 (0.39) | 1.25 (1.06) | 1.17 (1.03) |
| 7;8–9;3 | 12 | 17 | 3.00 (0.95) | 3.00 (1.12) | 1.67 (0.89) | 2.18 (1.07) | 2.92 (0.90) | 2.65 (1.46) |
| 9;4–10;11 | 7 | 22 | 2.57 (1.27) | 3.00 (1.02) | 1.86 (0.90) | 1.95 (0.90) | 2.71 (1.60) | 3.36 (1.43) |
| all | 35 | 51 | 2.54 (1.12) | 2.78 (1.14) | 1.43 (0.74) | 1.84 (0.95) | 2.11 (1.37) | 2.61 (1.59) |

*Note*. Possible range of performance for Recursive ToM Understanding task: 0–5, for Recursive ToM Production: 0–3, for Syntactic Recursion task: 0–4.

https://doi.org/10.1371/journal.pone.0266959.t005

**Table 6. Results of the logistic regression model predicting children's TB performance with their age in months and their performance in tasks of syntactic recursion (Synt. Recurs.), Recursive ToM understanding (RToM U), and Recursive ToM Production (RToM P).**

| | | | | 95% CI for Odds Ratio | | |
|---|---|---|---|---|---|---|
| | *B(SE)* | *z* | *p* | Lower | Odds Ratio | Upper |
| Included | | | | | | |
| Constant | -3.09 (1.52) | -1.88 | .04 | | | |
| Age in months | 0.02 (0.02) | 1.04 | .15 | 0.99 | 1.02 | 1.06 |
| Synt. Recurs. | 0.12 (0.24) | 0.41 | .62 | 0.70 | 1.12 | 1.80 |
| RToM P | 0.26 (0.36) | 0.73 | .47 | 0.66 | 1.30 | 2.75 |
| RToM U | 0.11 (0.20) | 0.54 | .59 | 0.75 | 1.11 | 1.67 |

*Note*. $R^2$ = .16 (Nagelkerke). Model $X^2(3)$ = 2.95, *p* = .40

https://doi.org/10.1371/journal.pone.0266959.t006

*Logistic regression models*. We again removed children failing the first-order FB condition (or having no FB trial available due to incorrect answers to the control questions in the FB condition, *n* = 10) from the following analyses. Prior to fitting the model, we checked for the assumptions. We checked for multicollinearity (all *VIF*s $\leq$ 1.49) and linearity of the logit for Age (*b* = 0.11, *p* = .61), recursive ToM understanding (*b* = 1.32, *p* = .43), recursive ToM production (*b* = 1.32, *p* = .43) and syntactic recursion (*b* = 1.93, *p* = .15).

None of the predictors in the full model predicted significantly children's TB performance (Table 6). The comparison of the full model with the null model containing the control variable only (TB ~ age) was not significant ($X^2(3)$ = 2.95, *p* = .40).

*Post-hoc comparisons*. We did not compute post-hoc one-sided two-sample Wilcoxon tests for TB-passers versus TB-non-passers in recursive ToM as it did not significantly predict the outcome in the logistic regression model in Study 2.

## Discussion

Study 2 did not show the relationship between children's TB performance and their recursive ToM found in Study 1, neither operationalized in terms of their understanding for recursive ToM nor in their own production of recursive ToM. The same task for children's recursive ToM understanding was used as in Study 1, however, the task was extended by an additional storyline and additional test questions and the task for recursive ToM production was added. This increased the duration of testing significantly and may have had a negative outcome on children's concentration and motivation in the test session and, therefore, the validity of results.

# Study 3

We therefore conducted Study 3 in which the initial relationship between children's TB performance and their recursive ToM understanding was tested without additional tasks testing for Advanced Pragmatics or recursive ToM production to reduce the duration of the test session and hold children's concentration as constant as possible.

## Method

**Participants.** The final sample included sixty-four 6- to 10-year-old children (72–131 months, *mean age* = 102 months; 32 girls, 32 boys). Five additional children were tested but excluded from data analyses because of technical issues during the test session (*N* = 4) and uncooperative behavior (*N* = 1).

**Material.** The task for syntactic recursion, the standard change-of-location task and the task for recursive ToM understanding were administered with the same material and procedure as in Study 2. In contrast to Study 2, there was no termination rule for the task testing for recursive ToM understanding.

## Results

**Coding of predictors.** As in Study 2, children were coded with the highest level of recursion up to which they answered all test question correctly (that is, one test question in the syntactic recursion task and two test questions in the recursive ToM understanding task).

**Plan of analysis.** The analysis was conducted in close similarity to Study 1 and 2. In scope of the preliminary analyses, we again tested for the typical performance in the TB and FB task in computing comparisons against chance level performance for both TB and FB in the three age groups (young, middle, old) and computed correlations between FB and TB performance.

To address the main research question of factors that influence the performance in the TB task, we again computed a logistic regression model. In the logistic regression model, TB performance (passing vs. no-passing) was predicted by children's age, recursive syntactic abilities and recursive ToM understanding. We compared this full model with a control model containing only children's age.

**Preliminary analyses: TB and FB performances in different age groups of children.** The performance in the change-of-location task as a function of belief type and age is depicted in Fig 6. Children were included when they answered the respective control questions correctly.

To test for the failure in the TB condition and the success in the FB condition of younger children and the success in both conditions in older children, we computed two-sided Wilcoxon tests against chance level performance (0.5) for the three age groups and the two belief

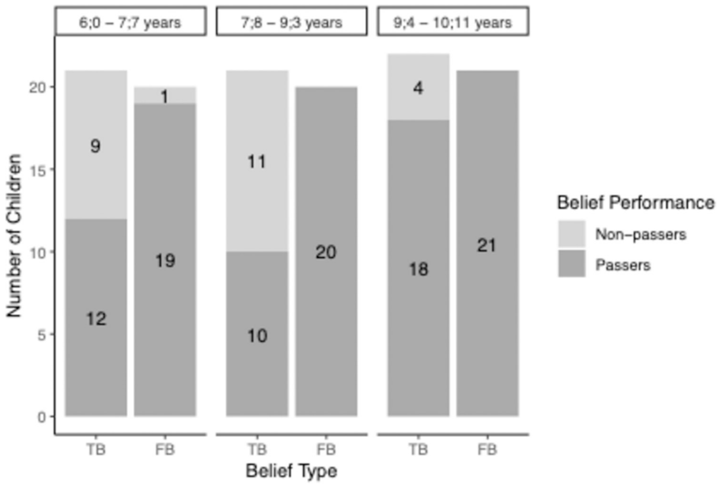



**Fig 6. Children's performance in the standard change-of-location task as a function of belief type and age in Study 3.** *Note*. Sample size of the groups vary as only children were included who answered the respective control questions correctly.

https://doi.org/10.1371/journal.pone.0266959.g006

**Table 7. Mean performance (M) and standard deviations (SD) in understanding of syntactic recursion (Synt Recurs.) and Recursive ToM Understanding (RToM_U) for TB non-passers and TB passers in three groups of age.**

| | *n* | | Synt. Recurs. | | RToM U | |
|---|---|---|---|---|---|---|
| | | | *M*(SD) | | *M*(SD) | |
| Age | noTB | TB | noTB | TB | noTB | TB |
| 6;0–7;7 | 9 | 12 | 2.11 (1.17) | 2.83 (0.83) | 2.00 (0.50) | 1.25 (0.62) |
| 7;8–9;3 | 11 | 10 | 2.55 (1.21) | 3.00 (0.94) | 1.73 (1.01) | 2.50 (1.08) |
| 9;4–10;11 | 4 | 18 | 3.00 (1.15) | 3.11 (1.28) | 2.75 (1.71) | 3.50 (1.42) |
| all | 24 | 40 | 2.46 (1.18) | 3.00 (1.06) | 2.00 (1.02) | 2.58 (1.48) |

*Note*. Possible range of performances for syntactic recursion task: 0–4, for Recursive ToM Understanding task: 0–5.

https://doi.org/10.1371/journal.pone.0266959.t007

conditions. The Wilcoxon tests showed that the youngest age group (6;0–7;7-year-olds) performed significantly above chance in the FB condition ($M = .95$, $p < .0001$, $r = -.87$). Wilcoxon tests could not be computed for the two older age groups due to ceiling effects in the FB condition (7;8–9;3-year-olds and 9;4–10;11-year-olds $M = 1$). In contrast, Wilcoxon tests revealed that in the TB condition, only the oldest age group of children (9;4–10;11-year-olds) performed significantly above chance ($M = .82$, $p < .01$, $r = -.65$). Younger children performed at chance level (6;0–7;7-year-olds: $M = .57$, $p = .53$, r = -.14, 7;8–9;2-year-olds: $M = .48$, $p = .84$, $r = -.04$).

The correlation between the TB and FB performance in the change-of-location task is (non-significantly) negative for the whole sample ($r$(phi) = -.10, $p = .46$) and (non-significantly) negative for the youngest age group (6;0–7;7-year-olds: $r$(phi) = -.19, $p = .43$). Because of the ceiling effects in the false belief condition, the correlation is not computable for the two older age groups.

**Main analyses: Predictors for TB performance.** *Descriptive statistics*. The mean performances in the syntactic recursion task and the recursive ToM understanding task as a function of TB performance (passers vs. non-passers) and age (young, middle, old age group) are summarized in Table 7.

*Logistic regression models*. We again removed children failing to answer first-order FB test and/or control questions correctly ($n = 4$) from the following analyses. Prior to fitting the model, we checked for the assumptions. We checked for multicollinearity (all *VIF*s $\leq 1.55$) and linearity of the logit for age ($b = 0.19$, $p = .40$), syntactic recursion ($b = -2.42$, $p = .11$) and recursive ToM understanding ($b = 1.61$, $p = .18$).

None of the predictors in the full model predicted significantly children's TB performance (Table 8). The comparison of the fit of the full model with the null model with the control variable only (TB ~ age) was not significant ($X^2(1) = 0.07$, $p = .79$).

**Table 8. Results of the logistic regression model predicting children's TB performance with their age in months and their performance in tasks of syntactic recursion (Synt. Recurs.) and Recursive ToM understanding (RToM U).**

| | | | | 95% CI for Odds Ratio | | |
|---|---|---|---|---|---|---|
| | *B(SE)* | *z* | *p* | Lower | Odds Ratio | Upper |
| Included | | | | | | |
| Constant | -2.80 (1.80) | -1.56 | .12 | | | |
| Age in months | 0.02 (0.02) | 1.23 | .26 | 0.98 | 1.02 | 1.06 |
| Synt R | 0.34 (0.26) | 1.31 | .19 | 0.85 | 1.40 | 2.37 |
| RToM U | 0.08 (0.29) | 0.26 | .79 | 0.61 | 1.08 | 1.93 |

Note. $R^2 = .12$ (Nagelkerke). Model $X^2(1) = 0.07$, $p = .13$.

https://doi.org/10.1371/journal.pone.0266959.t008

**Table 9. Partial correlation of predictors in Study 1–3 controlled for children's age in months.**

| | Study 1 | | | | | Study 2 | | Study 3 |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 1 |
| 1. Synt | | | | | | | | |
| 2. RToM P | | | | | | .27** | | |
| 3. RToM U | .30** | | | | | .07 | .11 | .32** |
| 4. Meta | .15 | | .15 | | | | | |
| 5. Irony1 | -.16 | | -.04 | -.14 | | | | |
| 6. Irony2 | .20 | | .00 | .01 | .37*** | | | |

** indicates $p < .01$

*** indicates $p < .001$.

https://doi.org/10.1371/journal.pone.0266959.t009

## Discussion

Similar to Study 2, the present study shows the typical TB performance in children between six and ten years but fails to replicate the relationship between children TB performance and their recursive ToM understanding that was found in Study 1. As neither children's syntactic recursion abilities nor their recursive ToM understanding predicted their TB performance, the results of the study therefore do not match with the predictions made by the pragmatic performance analysis.

**Supplementary explorative analysis: Correlations of predictors in Study 1–3.** In addition to the planed analysis, we conducted an exploratory analysis to investigate how the various predictors relate to each other. To this end, we computed partial correlations for the predictors in each study controlled for children's age in months (Table 9).

Table 9 shows a medium correlation between children's performance in the syntactic recursion task and recursive ToM understanding task for Study 1 and 3, but not in study 2. In study 2, however, children's performance in the syntactic recursion and recursive ToM production task correlated significantly. Additionally, children's performance in the first and second irony test question of the advanced pragmatics task in Study 1 show a medium to high correlation (Table 9).

## General discussion

Background to the present study was a puzzling empirical phenomenon: studies that administered the TB version of the classical change-of-location task to a broad age range of children yielded a surprising U-shaped performance curve: young children master this task, then from around age four children come to fail and performance only recovers around age eight to ten. Based on evidence that suggests that the decline in performance around age four reflects pragmatic confusions, the current set of studies tested whether performance recovery at the end of the U-shaped performance curve can be explained by another developmental change in children's pragmatic understanding. The studies therefore aimed to replicate the typical TB pattern (in an online testing format) and addressed potential factors that might explain the end of the U-shaped curve in two ways:

1. Is the TB pattern a function of advanced pragmatic development?
2. Even more generally: Is it a function of recursive ToM or recursive thinking in general?

The results, first of all, replicate the typical TB pattern in children between six and ten years: only the oldest age group (9;4–10;11 years) answered the TB task correctly while younger children performed at chance level. Investigating potential factors for the TB performance, the studies extended existing research methods by measuring children's recursive ToM and syntactic recursive abilities. Results revealed that children's syntactic recursion abilities and their recursive ToM understanding (Study 1 and 3) and recursive ToM production (Study 2) correlated substantially–indicating that they tap a common underlying capacity for recursive thinking. However, the results regarding the main research question remain inconclusive: Study 1 suggests that children's recursive ToM, but not their advanced pragmatics understanding or general recursive thinking abilities, predict their TB performance. This relationship, though, could not be replicated in Studies 2 and 3 in which neither recursive ToM nor recursive thinking in general explained children's performance in the TB task.

The studies overall, therefore, do not provide clear evidence for the pragmatic analyses of children's performance recovery in the TB task around age eight to ten. Of course, absence of evidence for a solid relationship between advanced pragmatics understanding, recursive ToM and recursion in general and TB performance does not amount to evidence of absence of any such relationship. There might be an actual relationship between children's advanced pragmatics and related factors and their TB performance that could not be reliably shown in the current set of studies because of different reasons. There are different possibilities regarding how this may be the case about which we can here only speculate.

One possibility is that there is an actual relationship, at least, between recursive ToM understanding and TB performance as shown in Study 1, that is indeed less pronounced than Study 1 indicated. The sample size calculations for the subsequent studies were based on the potentially overestimated medium effect size from Study 1 which possibly caused Studies 2 and 3 to be too underpowered to detect an effect that may be more subtle. Future studies need to address this potential issue with adequate sample sizes that would also detect smaller relationships.

Another possibility is that we failed to detect an actual relationship because of the implementation of the various predictors, especially children's advanced pragmatic understanding. This was operationalized as children's understanding for indirect speech acts, more specifically, their understanding of metaphors and irony–based on approaches that suppose non-literal language comprehension involves the ascription of complex intentions. The pragmatic analysis of the TB tasks predicts that, by ascribing such complex communicative intentions, children at the end of the U-shape curve resolve their pragmatic confusions. However, much daily non-literal language including frozen metaphors (whose metaphorical content is "dead") and frequently used ironic remarks is conventionalized in language [24, 41]. In cases of such conventionalized metaphor use as "The ATM swallowed my credit card" the hearer usually understands what is said without processing the literal meaning first and making inferences about the speaker meaning in a second step [42 p. 116]. Similarly, ironic utterances that are used with high frequency and familiarity (e.g., "That's just great") might become, to some degree, conventionalized. As a consequence, their non-literal meaning might become directly accessible [41], (*see also* [43] for conventionalized versus non-conventionalized indirect requests). In contrast, understanding un-conventionalized non-literal language requires the hearer to make pragmatic inferences based on her world knowledge, the context and the lexical meaning of the utterance [24 p. 247] and therefore tap a different set of cognitive skills than conventionalized non-literal language does [44].

Another possibility is that we failed to detect a relation to children's non-literal language understanding since the relevant pragmatic abilities involved in the comprehension of the TB task, on the one hand, and of irony and metaphor comprehension, on the other hand, might be quite different. (We thank the anonymous Reviewer of PloS ONE for sharing this issue in their review of an earlier version of this manuscript). Ironic and metaphoric speech acts involve a mismatch of sentence and speaker meaning. The pragmatic challenge is that the hearer needs to overcome the literal sentence meaning to be able to grasp the speaker meaning. In the TB task, however, sentence and speaker meaning are aligned. The challenge of the TB task might rather be to understand the speaker's conversational goal as the task and the answer to the test question are obvious and common ground and do not allow any other perspective on the scenario. Both the understanding for non-literal language and the TB task might be applications of advanced pragmatics understanding. However, they might not necessarily be seen as a unified phenomenon nor be subject to a same development [45].

Future studies, therefore, need to carefully operationalize these predictors of advanced pragmatics in order to measure children's abilities to ascribe recursive communicative intentions in valid ways. To this end, future work will need to test whether an operationalization with non-conventionalized non-literal language (such as novel metaphors and unfamiliar ironic utterances) or other pragmatics measures (e.g., understanding academic test question in general, understanding of literal meaning that changes with variations in context etc.) may succeed in measuring children's abilities for pragmatic inferences and recursive intention ascription and finding a relationship to children's TB performance.

Further future studies might test children's abilities in ascribing recursive intentions beyond the scope of (non-literal) verbal language understanding. An alternative that completely avoids any problems of interference of individual language knowledge would be to measure children's abilities in ascribing recursive communicative intentions in non-verbal scenarios. That might test the capacity for recursive, higher-order mindreading (that may also be applied to pragmatic language use) more generally. Cooperative coordination scenarios such as, for example, *Stag Hunt* scenarios, provide an elegant opportunity to ask children to reason recursively about other's non-verbal communicative acts. Stag Hunt scenarios [based on a parable by Jean-Jacque Rousseau, see 46] are game-theoretic interactions that require participants to cooperate with their fellow players to achieve a joint goal. Typically, they represent situations in which a player can choose between two options: the player can either hunt a *hare* (i.e. win a low value price) on her own or hunt a *stag* (i.e. win a high value price). If she decides to hunt a hare, the player will succeed independently of the other player's decision; if she decides to hunt a stag, in contrast, she only succeeds if the other player does so as well [46]. In a child-friendly adaptation, Wyman and colleagues showed that preschoolers succeed in coordinating with an adult co-player based on minimal non-verbal communication [47]. As the high-value option was only occasionally available, children needed to base their decision not only on what they themselves saw ("There is a stag") but also on what they thought about their fellow player ("She saw that there is a stag") and on what they inferred what their fellow player potentially thought about the them ("She knows that I know that she knows that there is a stag"). Children inferred the mutual knowledge of what their fellow player saw, knew and intended based on minimal non-verbal communicational cues (eye contact and smiling) and successfully based their decision for one of the two options on these inferences [47]. In a structurally related task by Grueneisen and colleagues, children were tested in a peer coordination scenario [48]. Children had to anticipate their partner's game decision based on a second-order mental state ascription ("She does not know that I know that *p*. She thinks that I falsely believe that *q*.") and had to coordinate their own game decision accordingly. Six-year-olds demonstrated their capacity to use such second-order mental state attributions to successfully coordinate with their peer without any communication [48]. Future studies could be based on such methods and thereby ask children to coordinate non-verbally in game scenarios with increasing complexity. Such adaptations might involve multiple co-players and, therefore, require even more complex recursive inferences (e.g., "Player 1 thinks that Player 2 did not see that Player 3 saw that there is a stag"). A systematic analysis of children's decisions in such game scenarios might then indicate the level of recursion at which they are able to reason.

A further line of future research will need to address the more general question of a unitary, domain-general development of recursive capacities. The current studies tested for children's recursive syntactic abilities until fourth order of recursion and children's recursive ToM understanding until fifth order of recursion as well as children's recursive ToM production in bluff scenarios. The results of correlated performance in recursion of mental representations and syntactic recursion provide first evidence for a shared underlying ability of general recursive thinking. In future research, these preliminary results need to be validated and extended over various forms of recursive thinking. Recently, it was theorized that embedding of recursive temporal representations shares conceptual similarities with embedding of recursive mental state representations [36]. Future studies thus need to compare children's development in holding recursive temporal representations (i.e. different forms of mental time travel) and higher-order ToM. Potential parallel trajectories of increasing levels of reasoning independently of its domain (i.e. mental states or temporal representations) would indicate shared underlying capacities of general recursive reasoning.

## Conclusion

In conclusion, the current set of studies replicate the typical TB performance pattern in children between six and ten years in an online testing format but do not provide clear evidence for an underlying development in advanced pragmatics that can explain this pattern. Future studies need to investigate more thoroughly whether this absence of evidence marks evidence of absence of any such relation between pragmatics and TB performance; or whether such relations exist but can be tapped only with suitably modified measures.

## Supporting information

**S1 File.**
(PDF)

**S1 Data.**
(XLSX)

**S2 Data.**
(XLSX)

**S3 Data.**
(XLSX)

## Acknowledgments

We thank Marlen Kaufmann for the organization of data collection, Lia Künnemann for the help with data collection and Cathrin Degen and Mellory Kripzak for help with reliability coding.

## Author Contributions

**Conceptualization:** Lydia Paulin Schidelko, Marina Proft, Hannes Rakoczy.

**Data curation:** Lydia Paulin Schidelko.

**Formal analysis:** Lydia Paulin Schidelko.

**Investigation:** Lydia Paulin Schidelko.

**Methodology:** Lydia Paulin Schidelko, Marina Proft, Hannes Rakoczy.

**Project administration:** Lydia Paulin Schidelko.

**Resources:** Hannes Rakoczy.

**Supervision:** Marina Proft, Hannes Rakoczy.

**Visualization:** Lydia Paulin Schidelko.

**Writing – original draft:** Lydia Paulin Schidelko.

**Writing – review & editing:** Marina Proft, Hannes Rakoczy.

## References


1. Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behav Brain Sci. Dezember 1978; 1(4):515–26.
2. Wimmer H, Perner J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition. 1983; 13(1):103–28. https://doi.org/10.1016/0010-0277(83)90004-5 PMID: 6681741
3. Wellman HM, Cross D, Watson J. Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. Child Dev. 2001; 72(3):655–84. https://doi.org/10.1111/1467-8624.00304 PMID: 11405571
4. Fabricius WV, Boyer TW, Weimer AA, Carroll K. True or false: Do 5-year-olds understand belief? Dev Psychol. 2010; 46(6):1402–16. https://doi.org/10.1037/a0017648 PMID: 21058830
5. Hedger JA, Fabricius WV. True Belief Belies False Belief: Recent Findings of Competence in Infants and Limitations in 5-Year-Olds, and Implications for Theory of Mind Development. Rev Philos Psychol. 2011; 2(3):429–47.
6. Oktay-Gür N, Rakoczy H. Children's difficulty with true belief tasks: Competence deficit or performance problem? Cognition. 2017; 166:28–41. https://doi.org/10.1016/j.cognition.2017.05.002 PMID: 28554083
7. Schidelko LP, Schünemann B, Rakoczy H, Proft M. Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task. Front Psychol. 2021; 12:4573. https://doi.org/10.3389/fpsyg.2021.703238 PMID: 34721151
8. Rakoczy H, Oktay-Gür N. Why Do Young Children Look so Smart and Older Children Look so Dumb on True Belief Control Tasks? An Investigation of Pragmatic Performance Factors. J Cogn Dev. 2020; 21 (2):213–39.
9. Searle JR. Speech acts: An essay on the philosophy of language. MA: Cambridge University Press; 1969. (Cambridge).
10. Siegal M. Language and thought: the fundamental significance of conversational awareness for cognitive development. Dev Sci. 1999; 2(1):1–14.
11. Siegal M. Knowing children: Experiments in conversation and cognition. Psychology Press; 2013.
12. Welch-Ross MK. Interviewer Knowledge and Preschoolers' Reasoning About Knowledge States Moderate Suggestibility. Cogn Dev. Juli 1999; 14(3):423–42.
13. Waterman AH, Blades M, Spencer C. Indicating when you do not know the answer: The effect of question format and interviewer knowledge on children's 'don't know' responses. Br J Dev Psychol. September 2004; 22(3):335–48.
14. Helming KA, Strickland B, Jacob P. Making sense of early false-belief understanding. Trends Cogn Sci. April 2014; 18(4):167–70. https://doi.org/10.1016/j.tics.2014.01.005 PMID: 24612994
15. Helming KA, Strickland B, Jacob P. Solving the Puzzle about Early Belief-Ascription. Mind Lang. 2016; 31(4):438–69.
16. Westra E. Pragmatic Development and the False Belief Task. Rev Philos Psychol. 2017; 8(2):235–57.
17. Siegal M, Beattie K. Where to look first for children's knowledge of false beliefs. Cognition. Januar 1991; 38(1):1–12. https://doi.org/10.1016/0010-0277(91)90020-5 PMID: 2015753
18. Papafragou A, Cassidy K, Gleitman L. When we think about thinking: The acquisition of belief verbs. Cognition. Oktober 2007; 105(1):125–65. https://doi.org/10.1016/j.cognition.2006.09.008 PMID: 17094956
19. Pouscoulous N, Tomasello M. Early birds: Metaphor understanding in 3-year-olds. J Pragmat. Januar 2020; 156:160–7.

**20.** Happé FGE. Communicative competence and theory of mind in autism: A test of relevance theory. Cognition. 1993; 48(2):101–19. https://doi.org/10.1016/0010-0277(93)90026-r PMID: 8243028

**21.** Winner E, Gardner H. Metaphor and irony: Two levels of understanding. Metaphor Thought. 1993; 2:425–43.

**22.** Sperber D, Wilson D. The Interpretation of Utterances: Semantics, Pragmatics & Rhetoric. 1986; 338.

**23.** Ariel M. Defining pragmatics. Cambridge University Press; 2010.

**24.** Matthews D. Pragmatic Development in First Language Acquisition. 2014. https://doi.org/10.1080/10489223.2014.962140 PMID: 26997850

**25.** Ackerman BP. Young Children's Understanding of a Speaker's Intentional Use of a False Utterance. Dev Psychol. 1981; 17(4):472.

**26.** Filippova E, Astington JW. Further development in social reasoning revealed in discourse irony understanding. Child Dev. 2008; 79(1):126–38. https://doi.org/10.1111/j.1467-8624.2007.01115.x PMID: 18269513

**27.** Filippova E. Irony Production and comprehension. In: Pragmatic Development in First Language Acquisition. John Benjamins Publishing Company; 2014. S. 261–78.

**28.** Lecce S, Ronchi L, Del Sette P, Bischetti L, Bambini V. Interpreting physical and mental metaphors: Is Theory of Mind associated with pragmatics in middle childhood? J Child Lang. März 2019; 46(2):393–407. https://doi.org/10.1017/S030500091800048X PMID: 30442207

**29.** Del Sette P, Bambini V, Bischetti L, Lecce S. Longitudinal associations between theory of mind and metaphor understanding during middle childhood. Cogn Dev. 56(100958).

**30.** Norbury CF. The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. Br J Dev Psychol. 2005; 23(3):383–99.

**31.** Angeleri R, Airenti G. The development of joke and irony understanding: A study with 3- to 6-year-old children. Can J Exp Psychol. 2014; 68(2):133–46. https://doi.org/10.1037/cep0000011 PMID: 24364812

**32.** Sullivan K, Zaitchik D, Tager-Flusberg H. Preschoolers Can Attribute Second-Order Beliefs. Dev Psychol. 1994; 30(3):395–402.

**33.** Liddle B, Nettle D. Higher-Order Theory of Mind and Social Competence in school-age Children. J Cult Evol Psychol. 2006; 4:215–29.

**34.** O'Grady C, Kliesch C, Smith K, Scott-Phillips T. The ease and extent of recursive mindreading, across implicit and explicit tasks. Evol Hum Behav. 2015; 144(4):744–63.

**35.** Corballis MC. The Recursive Mind: the Origins of Human Language, Thought, and Civilization [Internet]. 2014 [zitiert 18. Oktober 2021]. Verfügbar unter: https://doi.org/10.1515/9781400851492

**36.** Gautam S, Suddendorf T, Henry JD, Redshaw J. A taxonomy of mental time travel and counterfactual thought: Insights from cognitive development. Behav Brain Res. November 2019; 374:112108. https://doi.org/10.1016/j.bbr.2019.112108 PMID: 31340172

**37.** Hauser MD, Chomsky N, Fitch WT. The faculty of language: What is it, who has it, and how did it evolve? Evol Hum Lang Biolinguistic Perspect. 2010; 298(November):14–42.

**38.** Pinker S, Jackendoff R. The faculty of language: What's special about it? Cognition. 2005; 95(2):201–36. https://doi.org/10.1016/j.cognition.2004.08.004 PMID: 15694646

**39.** Arslan B, Hohenberger A, Verbrugge R. Syntactic recursion facilitates and working memory predicts recursive theory of mind. PLoS ONE. 2017; 12(1):1–23. https://doi.org/10.1371/journal.pone.0169510 PMID: 28072823

**40.** Happé FGE. An Advanced Test of Theory of Mind: Understanding of Story Characters' Thoughts and Feelings by Able Autistic, Mentally Handicapped, and Normal Children and Adults 1. J Autism Dev Disord. 1994; 24(2).

**41.** Burnett DL. Exploring the role of conventionality in children's interpretation of ironic remarks. J Child Lang. November 2015; 42(6):1267–88. https://doi.org/10.1017/S0305000914000798 PMID: 25499699

**42.** Recanati F. Literal meaning. Cambridge University Press; 2004.

**43.** Deliens G, Papastamou F, Ruytenbeek N, Geelhand P, Kissine M. Selective Pragmatic Impairment in Autism Spectrum Disorder: Indirect Requests Versus Irony. J Autism Dev Disord. September 2018; 48(9):2938–52. https://doi.org/10.1007/s10803-018-3561-6 PMID: 29633109

**44.** Pouscoulous N. Metaphor: For adults only? Belg J Linguist. 5. Dezember 2011; 25:51–79.

**45.** Geurts B, Kissine M, van Tiel B. Pragmatic reasoning in autism. In 2019.

**46.** Skyrms B. The Stag Hunt. Proc Addresses Am Philos Assoc. 2001; 75(2):31–41.

47. Wyman E, Rakoczy H, Tomasello M. Non-verbal communication enables children's coordination in a „Stag Hunt"game. Eur J Dev Psychol. 2013; 10(5):597–610.

48. Grueneisen S, Wyman E, Tomasello M. „I Know You Don't Know I Know. . ."Children Use Second-Order False-Belief Reasoning for Peer Coordination. Child Dev. 2015; 86(1):287–93. https://doi.org/10.1111/cdev.12264 PMID: 25040465

**Appendix D: Schidelko, Baumann, Proft, Rakoczy (2023)**

# Do Theory of Mind and Mental Time Travel abilities build on joint cognitive foundations? A Comparison of development of recursive Theory of Mind and Mental Time Travel

Lydia Paulin Schidelko, Leonie Baumann, Marina Proft, & Hannes Rakoczy

*Department of Developmental Psychology, University of Göttingen, Göttingen, Germany*

[Unpublished Manuscript]

**Abstract**

The developmental transition children undergo around age four when they come to master verbal false belief (FB) tasks is considered as onset of a meta-representational Theory of Mind (ToM) (Perner, 1991). Success in FB tasks goes along with emerging competence in conceptually related tasks (e.g., false sign tasks) that all require meta-representational perspective taking (e.g., Perner et al., 2002, 2003). The present study aims to elucidate the emergence and development of meta-representation more generally and to test claims of a cognitive foundation in growing meta-representation thought that is shared with other cognitive abilities, such as mental time-travel MTT (Gautam et al., 2019).

To this end, developing meta-representational ToM in three- to eight-year-old children's (*N*=120) was compared their Mental Time Travel (MTT) abilities on three levels of recursion. ToM was operationalized as first-, second-, and third-order FB understanding. MTT abilities were measured in children's understanding of future possibilities (first order), counterfactual thinking (second order) and anticipated counterfactual thinking (third order MTT).

The results show a stepwise development of first-, second- and third-order tasks in both ToM and MTT separately and a moderate consistency of performance pattern across both abilities. However, separate correlations of first-, second-, and third-order tasks of ToM and MTT showed only a relationship of second-order mental state ascription and counterfactual reasoning.

Overall, these results from a first direct comparison of first- to third-order ToM and MTT abilities confirmed evidence from earlier studies investigating both abilities separately in that development showed stepwise and parallel trajectories in both domains. However, the study does not provide conclusive results for actual associations of the two abilities and thus, no stringent evidence for shared foundations in meta-representational reasoning. Future studies need to address the question how different forms of recursive, meta-representational thinking can be measured validly in order to reveal developmental trajectories of meta-representational thought across domains.

One hallmark of human cognition is that it goes beyond representing the here and now. We do not only represent how we perceive the world from our own perspective, but also how it appears to agents – what others see, believe and desire (*Theory of Mind*). And, we do not only think about how the world is now, but also imagine other times – how the world was, could be or will be (*Mental Time Travel*). The present paper examines how these two ways of going beyond representations of the here and now are related in development: Do they share a joint cognitive foundation? Do they emerge together? And do they progress developmentally in parallel and coordinated ways?

Theory of Mind (ToM) is the social-cognitive ability to ascribe and reason about mental states (Premack & Woodruff, 1978). At the conceptual heart of ToM lies metarepresentation: the capacity to represent how others represent the world even when these representations diverge from the interpreter's own perspective, and from reality. Meta-representational ToM emerges in the course of a major conceptual transition around age four (Perner, 1991; Perner & Roessler, 2012). Around this time, children come to master the litmus test for meta-representation, the so-called False Belief (FB) task (in which they need to track a protagonist's subjective and outdated belief) as well as many other conceptually related tasks that differ with regard to topic, format and surface structure but all require meta-representational thinking (for an overview, *see* Perner & Roessler, 2012)

The second potentially unique hallmark of human cognition is complex Mental Time Travel (MTT) – the ability to think and reason about the past (episodic memory) and the future (foresight) and how they relate to the present (Bieri, 1986; Bischof-Köhler, 1985; Buckner & Carroll, 2007; Redshaw & Suddendorf, 2020; Schacter et al., 2007; Suddendorf & Corballis, 2007; Tulving, 1985).

Developmental research on the emergence of this capacity has focused on future-oriented forms of MTT. In prototypical tasks, participants are faced with a situation that opens up multiple, still undetermined future possibilities. Crucially, the uncertain future possibilities are mutually exclusive – such that only one of the possibilities will *actually* occur. This requires the participant to represent that and how these future possibilities relate to the same present state.

In a task designed by Beck and colleagues, for example, children were asked to put out mats to catch a mouse that will come out of either of two sides of a forked-shaped (upside-down y-shaped) slide. The results show that only children aged five and older succeed to prepare for this uncertain future event by putting two mats under the two exits of the forked-shaped slide (Beck et al., 2006). A simplified version of this task (Redshaw & Suddendorf, 2020) in which participants were allowed to cover the exits with their hands showed success in slightly younger children around age four. In another version of this task, two objects come down two slides simultaneously. One of the slides is non-branching (with one exit at the bottom), the other slide (the forked slide) branches into two exits at the bottom. Participants are asked to cover only one exit with a wagon in order to catch one of the two objects. In this set-up, they had to reflect on all options simultaneously in order to choose the certain over two uncertain options. Children aged five and older succeeded by choosing the certain option (i.e. put wagon under the non-branching slide) over the uncertain options (the exits of the branching slide) (Leahy & Zalnieriunas, 2021). Other studies which followed related task structures led to similar results (Robinson et al., 2006).

Children thus come to demonstrate MTT (reasoning about future possibilities) from around age four to five (Beck et al., 2006; Redshaw & Suddendorf, 2016; Robinson et al., 2006). Similarly, children at the age of four reliably come to solve FB and conceptually related tasks that require meta-representational ToM (Wellman et al., 2001). Are these parallel developmental trajectories a mere coincidence? Or could they reflect deep and underlying cognitive commonalities and foundations in the two capacities (Gautam et al., 2019)?

One possibility is that both ToM and MTT build on a joint (neurocognitive) capacity for simulation and projection as common cognitive foundations. Explicit ToM requires simulation to shift perception from the immediate environment to alternative perspectives. In a similar sense, imagining future events requires projection to shift perception from the immediate environment to the alternative, imagined future event (Buckner & Carroll, 2007). Another (not necessarily mutually exclusive) possibility is that both ToM and MTT not only share overlapping processes of projection and simulation in general but build specifically on the same underlying cognitive ability to hold meta-representations (Gautam et al., 2019; Redshaw, 2014; Tulving, 2005).

In case of (explicit) ToM, it is obvious that meta-representation in some sense underlies the capacity to represent the subjective perspectives (representations) of agents (Perner, 1991; Pylyshyn, 1978; Suddendorf, 1999). It is prima facie somewhat less obvious in which ways MTT may build on meta-representation. Different arguments have been made to this effect. One is that MTT requires a particular form of (autonoetic) representation of one's own past or potential future experiences and perspectives and thus a form of meta-representation (Tulving, 2005; see also Perner et al., 2007; Perner & Ruffman, 1995). Another one is that meta-representation is necessary for representing how various temporal perspectives relate to what they represent: What distinguishes remembering that *p*, or predicting that *p* from merely idly imagining that *p* is that the former two involve specific representational relations of the content entertained (*p*) to present or future states of affairs in the world; and so to understand that and how one can travel in time into the past or future, one has to have some minimal meta-representational grasp of the way temporal representations relate to the world (Bieri, 1986; Redshaw, 2014).

Against this background, the first research question of the present project is: Do the two ways of going beyond representations of the here and now under consideration – ToM and MTT – build on some form of joint (neuro-)cognitive foundation and therefore emerge in tandem in development (Gautam et al., 2019)?

As reviewed above, a big body of indirect evidence that is compatible with such a picture comes from studies that suggest roughly the same age of emergence for the two capacities (Beck et al., 2006; Leahy & Zalnieriunas, 2021; Redshaw & Suddendorf, 2016; Robinson et al., 2006; Wellman et al., 2001). But similar age of emergence does not present stringent evidence for shared foundations and related development. More direct evidence that the two types of tasks are correlated would be needed. Preliminary evidence along such lines comes from a recent study in which children were tested in both a modified version of the slides task (Beck et al., 2006; Leahy & Zalnieriunas, 2021) and a FB task (Perner et al., 1987). Results revealed no difference in the age of success in the ToM and future possibilities tasks (Halberda et al., 2023; but see Immel et al., 2022). Though both come to be mastered around the same age, more systematic and direct comparisons from comprehensive correlational studies are needed, and one rationale of the present study is to supply such evidence.

If the emergence of ToM and MTT actually show developmental associations pointing to joint underlying cognitive foundations of meta-representational thought, the following question arises: does this developmental association only rely on the joint emergence of meta-representational thought and, based on that, first-order ToM understanding and future MTT or is also subsequent development of more complex, higher-order forms of ToM and MTT related?

In subsequent development in both areas, more complex and higher-order forms of reasoning emerge by recursive embedding. In ToM, meta-representation can be potentially iterated ad infinitum ("A thinks that B thinks that C thinks … that *p*"). Children ascribe first-order beliefs around age four, and only after some developmental delay, attribute second-order mental states from around age five to six ("A thinks that B believes that. . .") (Sullivan et al., 1994). Very little is known about children's development of higher-order ToM beyond second-order of recursion. First evidence shows that children become able to attribute third- and fourth-order mental states only during middle childhood (Liddle & Nettle, 2006; Schidelko et al., 2022).

Similarly, recursive iteration can generate more complex and higher-order forms of MTT in which additional levels of temporal representations can be represented: remembering past moments in which one thought about earlier times or imagining that in the future one will make plans for the further future. Recursive temporal embedding can also involve switching back and forth between future and past perspectives: remembering moments in the past in which one thought about the future or imagining that in the future one will look back on the relative past (Redshaw & Suddendorf, 2020). One example of second-order recursive MTT is counterfactual reasoning about past and present events (like in "What would the US look like today if Trump had won the last election?"). Here one has to mentally go back in time to imagine how an alternative past (level 1) would have affected the future and thus led to an alternative present (level 2). Crucially, in this process, the relation between several simultaneous representations of the world (actual reality and counterfactual alternative) need to be represented (Beck & Guthrie, 2011; Byrne, 2016; McCormack et al., 2018; Rafetseder et al., 2013). Actual reality and the counterfactual alternative need to be represented as alternative versions of the very same moment in time that originate from a common past (Beck, 2016; Redshaw & Suddendorf, 2020).

To validly investigate counterfactual reasoning in development, tasks need to follow a stringent structure of double determination (e.g., Rafetseder et al., 2013, 2021). In these scenarios, the outcome is determined by at least two causes (e.g., a toy pic is hit by two rolling discs and thus falls down) but the counterfactual scenario eliminates only one of the causes (one disc is prevented from rolling against the toy pig) so that the outcome (toy pig falls down) stays the same (McCormack et al., 2018). The use of doubly-determination reduces the possibility that participants solve the counterfactual test question (e.g., "If I had not rolled the red disc down that time, would the pig have fallen down?") with simpler conditional reasoning (e.g., hypothetical thinking or if-then-principles) (e.g., Rafetseder et al., 2013, 2021). In this task, children from age six were able to answer counterfactual test questions about doubly-determined scenarios correctly (McCormack et al., 2018).

Around the same time, children typically also begin to experience counterfactual emotions such as regret and relief (McCormack et al., 2019; McCormack & Feeney, 2015; O'Connor et al., 2012, 2015; Uprichard & McCormack, 2019; Van Duijvenvoorde et al., 2014; Weisberg & Beck, 2010). These emotions build on second-order MTT in the following way: experiencing regret or relief requires the agent to realize that in the past (level 1) different options of the relative future (level 2) were available (Hoerl & McCormack, 2019). By comparing the true present and an alternative present, the actual state of the world is evaluated as better (relief) or worse (regret) than the imagined counterfactual.

Adding yet another layer of recursion, anticipation of counterfactual emotions constitutes third-order MTT and seems to emerge later in development than the experience of relief and regret itself. Only from around age eight or later, children report negative emotions when imagining that they will learn in the future (level 1) that – in the relative past (level 2) – better options of the relative future (level 3) were available (Guttentag & Ferrell, 2008; McCormack & Feeney, 2015; O'Connor et al., 2015; Rafetseder & Perner, 2012). Against this background, a second research question is: does subsequent development in both MTT and ToM rely on the same capacity to recursively embed representations and therefore develop in a parallel and correlated fashion?

To address the two research questions of joint underlying foundation of the emergence and subsequent development of ToM and MTT, we compared children's performance in first- to third-order ToM and first- to third-order MTT tasks.

## Methods

### Design

Each child was tested in all six tasks (*see* Table 1) by one of three female experimenters (E) in two separate test sessions. In test session A, children received all three ToM tasks, and in test session B they received all three MTT tasks. The order of sessions was counterbalanced. All children participated remotely in a video conference (following the procedure by Schidelko et al., 2021). For more details on the tasks, procedure and counterbalancing, *see* Supplement.

Table 1. Conceptualization and operationalization of first-, second- and third-order Theory of Mind tasks and first-, second- and third-order Mental Time Travel tasks.

| **Order** | **Theory of Mind** | | **Mental Time Travel** | |
|---|---|---|---|---|
| 1 | First-order False Belief | A thinks that p | Future Possibilities | Future A or Future |
| 2 | Second-order False Belief | B thinks that p<br>A thinks that | Counterfactual reasoning | Alternative Future<br>Past |
| 3 | Third-order False Belief | C thinks that p<br>B thinks that<br>A thinks that | Anticipation of regret | Alternative Present<br>Past<br>Future |

**Participants**

The final sample comprised one-hundred-and-twenty 3- to 8-year-old children (36 - 107 months, *M* = 71.51 months; 61 female, 59 male) consisting of twenty children per age group. Twenty additional children were tested but excluded from data analyses (for details on the sample and reasons for exclusion, *see* Supplement).

**Material**

**Mental Time Travel tasks**

**First-order Mental Time Travel Task**

The task tested for children's abilities to reason about future possibilities (adapted from Beck et al., 2006). In two test trials, children were asked to help a monkey to catch a banana ("In which hole should the monkey put the bucket, so that he catches a banana for sure?"; *see* Figure 1). Children answered by naming the color of the hole. The answer was rated as correct when children chose the single hole (e.g., *in Figure 1*: the green hole).


**Figure 1**. Screenshot of first-order MTT task with arrows indicating where bananas will fall.

**Second-order Mental Time Travel Task**

The task (adapted from McCormack et al., 2018) tested for children's counterfactual reasoning abilities with two test questions. Children saw a video sequence of two snowballs running down two hills breaking a snowman at the valley between the hills before E asked the test questions in counterbalanced order:

- Test Question A (subtractive counterfactual): "If the white snowball had not rolled down the hill this time, would the snowman then be broken?" [correct answer: yes]
- Test Question B (additive counterfactual): "If there had been the tree on the steep hill this time (*animated finger pointed at the steep hill),* would the snowman then be broken?" [correct answer: yes]

**Third-order Mental Time Travel Task**

The task tested if children anticipated that they would feel regret about a choice if they would learn that there was a better option in the past (McCormack & Feeney, 2015). To this end, children received two trials: a control trial and an experimental trial. In the both trials, children were asked to choose one of two boxes to win its content. Children always won one coin from their chosen box. The other box remained closed, and the children did not see what was actually inside. Children were then asked to answer the baseline question and the test question. The test question differed between the control and experimental trail (see below).

Baseline question: "How do you feel now that there was one coin in your box?"

Children rated their emotion on the smiley scale (Figure 3a). The emotion they chose became the anchor for the subsequent test question.

Test question: "This is how you felt, when one coin was in your box. Imagine, that there were five coins [experimental trial]/ one coin [control trial] in the other box, and you did not win them/it. How would you now feel about winning this box with one coin? Sadder, happier or the same?" (with 'this box' referring to the box they chose), see Figure 3b and c).

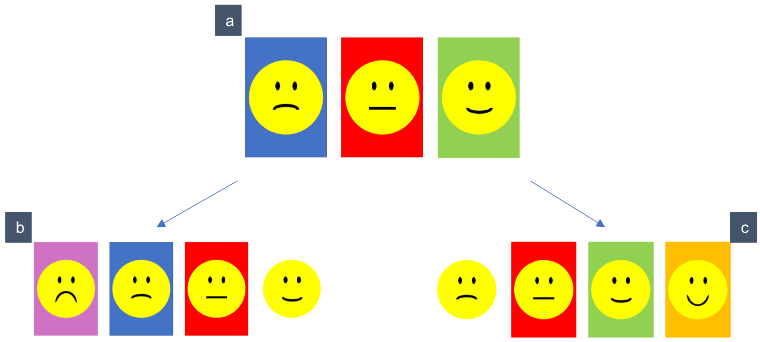



**Figure 3**. Smiley scale used in third-order MTT task. Example: Figure 4a: Baseline question: "How do you feel now that there was one coin in your box? Happy, sad or not happy and not sad", if child answers: "Happy" [Figure 4a → Figure 4c], if child answers: "Sad" [Figure 4a → Figure 4b], if child answers: "not happy and not sad" [Figure 4a → Figure 4a] for subsequent test question.

Children received scores for their baseline rating and their test question rating in both the control trial and the experimental trial. From the two scores within each trial (baseline versus test question), we computed two difference scores. The difference score from the experimental trial was then compared to the difference score of the baseline trial. Children were classified as being able to anticipate regret if they rated their emotional change in the experimental trial more negatively than in the control trial. Thus, to pass this task, the difference score of the experimental trial needed to be smaller than the baseline trial (for more information and examples of coding, *see* Supplement).

**Theory of Mind tasks**

Children saw three story lines that tested for their first-order FB understanding (two test questions), their second-order FB understanding (two test questions) and their third-order FB understanding (one test question).

**Example First-order ToM**

In this version of the standard-change-of-location FB task (Wimmer & Perner, 1983), the child saw that Protagonist A placed his object O (a ball) in one of two boxes (box 1). Protagonist A then left the scene and Protagonist B entered the scene, relocated O to box 2 and left again. After that, the following test question was asked.

- First-order FB Test Question 1: “When A comes back, where will A look for O first?” [correct answer: box 1]

**Example Second-order and third-order ToM**

Children listened to the video “The school soccer team” (Liddle & Nettle, 2006) and answered the following questions:

- Second-order FB Test Question 2: Which sentence is true?
  A) Max does not know that the soccer coach wants Max and Paul to both play on the soccer team.* [correct answer]
  B) Max knows that the soccer coach wants Max and Paul to both play on the soccer team.*
- Third-order FB Test Question: Which sentence is true?
  A) The soccer coach thinks that Max believes that he wants that Max to play on the soccer team.*
  B) The soccer coach thinks that Max believes that he wants that Max does not play on the soccer team.* [correct answer]

*German translation with that-complement for want (“möchte, dass”)

## Results

### Coding

Correct trials were coded with “1”, incorrect trials with “0”. All trials of a task were added to a sum score. In the first- and second-order tasks, children were able to receive a sum score between 0 and 2. In the third-order task, in contrast, children received only one trial per task and could therefore score 0 or 1. To be categorized as passers (indicated by a “+” in Table 2), children needed to solve all trials of a task correctly (i.e., two trials in the first- and second-order tasks and one trial in the third-order tasks). Children’s performance across three tasks of one ability (ToM or MTT) was categorized in performance patterns. We expected children to perform in one of four specific performance patterns (see Table 2). Performance that deviated from these patterns was categorized as “other”.

Table 2. Pattern of success (+) and failure (-). Performance that deviated from these patterns was categorized as “other”.

| | Pattern A | Pattern B | Pattern C | Pattern D |
|---|---|---|---|---|
| First-order Task | – | + | + | + |
| Second-order Task | – | – | + | + |
| Third-order Task | – | – | – | + |

### Plan of Analysis

In a first step, we aimed to test whether the development of both ToM and MTT abilities follow a stepwise order. To ensure that this sequence indicates progressive development, we ran Guttman scaling analyses on ToM and MTT separately (Wellman & Liu, 2004). We expected children to perform in one of four performance patterns, see Table 2. To analyze the reliability of the ToM and MTT Guttman scales, coefficients of reproducibility and coefficients of scalability were computed.

In a second step, we compared whether children’s performance pattern in ToM aligned with their performance pattern in MTT. To test the consistency of performance, we computed a Weighted Kappa of the cross table of the four expected performance patterns. Additionally, correlations of the sum scores between first-order, second-order and third-order tasks were calculated controlling for children’s age in months.

**Stepwise Development in Theory of Mind and Mental Time Travel**

The proportion of children who passed each task is presented in Table 3 for MTT tasks and in Table 4 for ToM tasks.

Table 3. Precents of children passing Mental time travel tasks across age groups.

| | 3-year-olds | 4-year-olds | 5-year-olds | 6-year-olds | 7-year-olds | 8-year-olds |
|---|---|---|---|---|---|---|
| MTT 1 | 30 | 50 | 70 | 75 | 90 | 95 |
| MTT 2 | 15 | 20 | 25 | 70 | 89 | 85 |
| MTT 3 | 5 | 10 | 25 | 35 | 25 | 65 |

Table 4. Precents of children passing ToM tasks across age groups.

| | 3-year-olds | 4-year-olds | 5-year-olds | 6-year-olds | 7-year-olds | 8-year-olds |
|---|---|---|---|---|---|---|
| ToM 1 | 35 | 95 | 85 | 85 | 95 | 95 |
| ToM 2 | 25 | 45 | 40 | 55 | 80 | 90 |
| ToM 3 | 35 | 30 | 20 | 20 | 40 | 65 |

In general, performance in both ToM and MTT followed a stepwise order (see Figure 5). Children first passed first-order tasks, then second- and finally third-order tasks. Performance of 83% of the children (100 of 120) fit this three task Guttman scale of ToM tasks, and performance of 85% of the children (102 of 120) fit the Guttman scale of MTT tasks. The coefficient of reproducibility (Green, 1956) for the scalogram analysis of ToM tasks was $C_R(TOM)$=.89 and $C_R(MTT)$=.90 for MTT tasks (Jobling & Snell, 1961). The Coefficient of Scalability is $C_S(ToM)$= 0.66 for ToM, and $C_S(MTT)$ = 0.72 for MTT (Menzel, 1953).

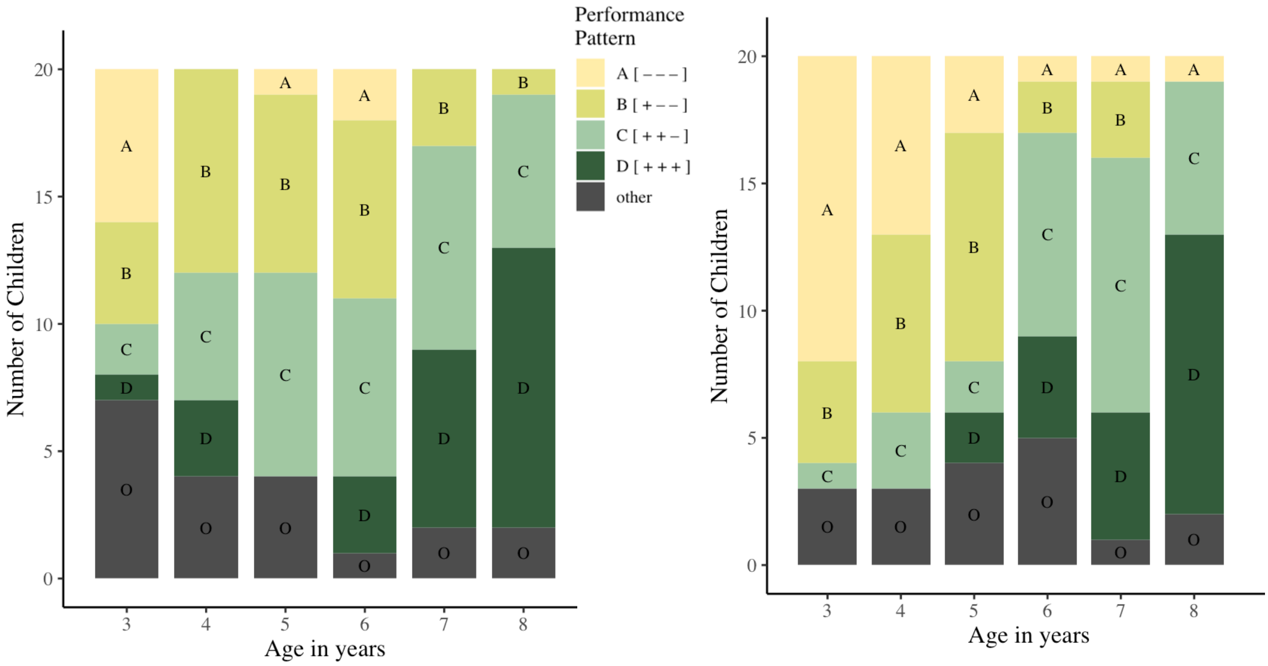



**Figure 5**. Number of children performing in each Performance pattern by function of their age in Theory of Mind (left) and Mental time Travel tasks (right).

### Consistency of Performance across Theory of Mind and Mental Time Travel

Table 5 shows children's performance patterns in MTT as a function of their performance pattern in ToM. The consistency of performance (without category "other") is moderate ($K = .31$). Correlations between the two first-, second- and third-order tasks respectively showed significant results only for the second-order tasks ($r = .26$, $p < .01$). First-order tasks ($r = -.03$, $p = .75$) and third-order tasks ($r = .11$, $p = .23$) did not correlate significantly when controlled for children's age.

Table 5. Number of children per performance pattern in Theory of Mind and Mental Time Travel.

| Pattern | MTT A | MTT B | MTT C | MTT D | MTT other |
|---|---|---|---|---|---|
| ToM A | 3 | 4 | 0 | 1 | 1 |
| ToM B | 8 | 9 | 6 | 2 | 5 |
| ToM C | 3 | 8 | 12 | 7 | 6 |
| ToM D | 3 | 1 | 9 | 10 | 2 |
| ToM other | 8 | 3 | 3 | 2 | 4 |

*Note.* Pattern A reflects failure across all task, pattern B reflects success only in first-order task, pattern C reflects success in first- and second-order task, pattern D reflects success across all tasks.

## General Discussion

The guiding question of the current study was whether two cognitive abilities – ToM and MTT – rely on joint cognitive foundations and thus, emergence and develop in parallel and correlated ways across childhood. To address this question, we aimed to investigate the emergence of both first-order ToM and MTT and subsequent development in higher-order forms of both ToM and MTT. The results show a stepwise development of first-, second- and third-order tasks in both ToM and MTT separately and a fairly consistent performance pattern across both abilities. However, separate correlations of first-, second-, and third-order tasks of ToM and MTT showed only a relationship of second-order mental state ascription and counterfactual reasoning.

To our knowledge, this study was the first attempt to directly compare children developing abilities in these two domains across different levels of recursive embedding. Overall, evidence from earlier studies investigating both abilities separately was confirmed in that we found parallel developmental trajectories in the two domains. However, the study does not provide conclusive evidence for actual associations of the two abilities and thus, no stringent evidence for shared foundations in meta-representational reasoning. Of course, absence of a solid relationship between ToM and MTT as evidence for joint underlying foundations does not amount to evidence of absence of any joint foundation. There might be an actual relationship between children's developing abilities in these two domains that could not be reliably shown in the study because of different reasons. There are different possibilities regarding how this may be the case about which we can here only speculate.

One possibility is that we failed to detect an actual relationship because of the methodological implementation of the tasks, especially children's MTT abilities. The three orders of MTT were operationalized as reasoning about future possibilities, counterfactual reasoning about past events, and anticipation of regret. However, our implementations of these tasks might have been artificially difficult as they imposed additional task demands. For example, the first-order MTT tasks testing for children's reasoning about future possibilities posed additional task demands in physical reasoning. In this task, children were asked to catch one of two bananas that were falling from trees. Performance in this task might heavily rely on children's physical reasoning abilities to imagine the physical trajectories of the falling bananas in addition to their MTT abilities. Future studies, therefore, need to carefully operationalize these concepts in order to measure children's MTT abilities in valid ways. To this end, future work will need to test whether operationalizations which do not rely on imaginations of physical trajectories may succeed in measuring children's abilities for reasoning about future possibilities and finding a relationship to children's first-order ToM. First support for the relevance of the operationalization in possibility reasoning tasks comes from a study that required reasoning about future possibilities based on object identities (Alderete & Xu, unpublished) rather then about locations based on physical trajectories. As in this task, even three-year-olds performed competently, performance diverged strikingly from performances in studies that used tasks that rely physical trajectories (Beck et al., 2006; Leahy & Zalnieriunas, 2021; Robinson et al., 2006).

Another possibility is that both ToM and MTT actually rely on meta-representational thought but we failed to detect a related performance as the proficiency and flexibility with which children can engage in meta-representational thought might vary with its content. Children's amounts of experience in the two domains might diverge in that they are more familiar with mental state reasoning than with MTT. Their life is full of social interactions with their parents, siblings, friends and teachers who actively motivate them to engage in mental state reasoning. In contrast, reasoning about other points in time might be less relevant in children's daily life. Children are not as frequently confronted with situation that require them to reason, for example, counterfactually. Indirect evidence for this claim comes from training studies that show that children's performance in ToM task benefits from practice in mental state reasoning

(Hofmann et al., 2016). Future studies will need to address this in comparing children's performance in less familiar meta-representational task implementation (as, for example, the false sign task that requires reasoning about false representation of a sign posts) with their emerging MTT abilities.

A further line of future research will need to address the more general question of the underlying developmental trajectory of recursive meta-representational thought. Different underlying developmental trajectories could account for the stepwise development of higher-order forms found for both ToM and MTT. One possibility is that each additional level of recursive embedding has to be learnt anew and thus, shows developmental delay which leads to a prolongend, stepwise development of higher-order meta-representational thought. Evidence for this claim would come from studies showing protracted, but parallel and correlated developmental progress and reasoning on the same level of embedding across domains. Another possibility is that higher-order meta-representational reasoning requires another qualitative change of the underlying thought system – another conceptual change in ToM development after the transition to first-order meta-representational thought (Miller, 2009; Perner, 1988). This developmental transition is needed for children to understand the recursive nature of mental states, that is, that mental states can be recursively embedded within each other. This principle of recursion in mental states needs to be understood to be then applied (Perner, 1988). Once children gain the insight to apply recursive operations on mental states, or, meta-representations, more generally, they are virtually able to hold meta-representations of any order of embedding. According to this idea, protracted development beyond second-order embeddings rather reflects performance limitation due to task factors or limited memory capacities but do not reflect competence limitation in recursive meta-representational thought itself. Future studies need to test this claim by measuring developing recursive abilities in different domains of meta-representational thought. To specifically address the hypothesis that task features (such as length of story line, unnecessary information) other than the level of complexity cause developmental delay after the cognitive transition to second-order ToM, task demands of higher-order ToM tasks should be reduced to a minimum. Additionally, higher-order ToM tasks should provide memory aids such as second-order FB tasks (Sullivan et al., 1994). Evidence of a more similar performance in such modified second- and higher-order task independent of the order of complexity would provide first evidence for the influence of task demands on the protracted performance beyond second-order reasoning.

**References**

Beck, S. R. (2016). Why What Is Counterfactual Really Matters: A Response to Weisberg and Gopnik (2013). *Cognitive Science*, *40*(1), 253–256. https://doi.org/10.1111/cogs.12235

Beck, S. R., & Guthrie, C. (2011). Almost thinking counterfactually: Children's understanding of close counterfactuals. *Child Development*, *82*(4), 1189–1198.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, *77*(2), 413–426. https://doi.org/10.1111/j.1467-8624.2006.00879.x

Bieri, P. (1986). Zeiterfahrung und Personalität. *Zeit, Natur Und Mensch. Beiträge von Wissenschaftlern Zum Thema „Zeit"*, 261–281.

Bischof-Köhler, D. (1985). Zur phyogenese menschlicher motivation. In *Emotion und reflexivitut (Emotion and reflexivity)* (pp. 3–47). Urban & Schwarzenberg.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49–57. https://doi.org/10.1016/j.tics.2006.11.004

Byrne, R. M. J. (2016). Counterfactual Thought. *Annual Review of Psychology*, *67*(1), 135–157. https://doi.org/10.1146/annurev-psych-122414-033249

Gautam, S., Suddendorf, T., Henry, J. D., & Redshaw, J. (2019). A taxonomy of mental time travel and counterfactual thought: Insights from cognitive development. *Behavioural Brain Research*, *374*, 112108. https://doi.org/10.1016/j.bbr.2019.112108

Green, B. F. (1956). A method of scalogram analysis using summary statistics. *Psychometrika*, *21*(1), 79–88. https://doi.org/10.1007/BF02289088

Guttentag, R., & Ferrell, J. (2008). Children's understanding of anticipatory regret and disappointment. *Cognition & Emotion*, *22*(5), 815–832. https://doi.org/10.1080/02699930701541542

Halberda, J., Cesana-Arlotti, N., Jauregui, S., & Nichols, S. (2023). *Developmental change in children's understanding of possible and impossible events in a "plinko" physical probability device*. Unpublished Manuscript.

Hoerl, C., & McCormack, T. (2019). Temporal updating, temporal reasoning, and the domain of time. *Behavioral and Brain Sciences*, *December*. https://doi.org/10.1017/S0140525X19001195

Hofmann, S. G., Doan, S. N., Sprung, M., Wilson, A., Ebesutani, C., Andrews, L. A., Curtiss, J., & Harris, P. L. (2016). Training children's theory-of-mind: A meta-analysis of controlled studies. *Cognition*, *150*, 200–212. https://doi.org/10.1016/j.cognition.2016.01.006

Immel, A.-S., Altgassen, M., Meyer, M., Endedijk, H. M., & Hunnius, S. (2022). Self-projection in early childhood: No evidence for a common underpinning of episodic memory, episodic future thinking, theory of mind, and spatial navigation. *Journal of Experimental Child Psychology*, *223*, 105481. https://doi.org/10.1016/j.jecp.2022.105481

Jobling, D., & Snell, E. J. (1961). The Use of the Coefficient of Reproducibility in Attitude Scaling. *The Incorporated Statistician*, *11*(2), 110. https://doi.org/10.2307/2987432

Leahy, B., & Zalnieriunas, E. (2021). *Might and might not: Children's conceptual development and the acquisition of modal verbs*. 20.

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

Liddle, B., & Nettle, D. (2006). Higher-Order Theory of Mind and Social Competence in school-age Children. *Journal of Cultural and Evolutionary Psychology*, *4*, 215–229. https://doi.org/10.1556/JCEP.4.2006.3

McCormack, T., & Feeney, A. (2015). The development of the experience and anticipation of regret. *Cognition & Emotion*, *29*(2), 266–280. https://doi.org/10.1080/02699931.2014.912201

McCormack, T., Ho, M., Gribben, C., O'Connor, E., & Hoerl, C. (2018). The Development of Counterfactual Reasoning about Doubly-Determined Events. *Cognitive Development*, *45*, 1–9.

McCormack, T., O'Connor, E., Cherry, J., Beck, S. R., & Feeney, A. (2019). Experiencing regret about a choice helps children learn to delay gratification. *Journal of Experimental Child Psychology*, *179*, 162–175. https://doi.org/10.1016/j.jecp.2018.11.005

Menzel, H. (1953). A New Coefficient for Scalogram Analysis. *Public Opinion Quarterly*, *17*(2), 268. https://doi.org/10.1086/266460

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological Bulletin*, *135*(5), 749–773. https://doi.org/10.1037/a0016854

O'Connor, E., McCormack, T., Beck, S. R., & Feeney, A. (2015). Regret and adaptive decision making in young children. *Journal of Experimental Child Psychology*, *135*, 86–92. https://doi.org/10.1016/j.jecp.2015.03.003

O'Connor, E., McCormack, T., & Feeney, A. (2012). The development of regret. *Journal of Experimental Child Psychology*, *111*(1), 120–127. https://doi.org/10.1016/j.jecp.2011.07.002

Perner, J. (1988). Developing semantics for theories of mind: From propositional attitudes. Developing theories of mind, 141. In *Developing theories of mind*. Cambridge University Press.

Perner, J. (1991). *Understanding the representational mind*. The MIT Press.

Perner, J., Kloo, D., & Gornik, E. (2007). Episodic memory development: Theory of mind is part of re-experiencing experienced events. *Infant and Child Development*, *16*(5), 471–490. https://doi.org/10.1002/icd.517

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137. https://doi.org/10.1111/j.2044-835X.1987.tb01048.x

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. (1995). Episodic Memory and Autonoetic Conciousness: Developmental Evidence and a Theory of Childhood Amnesia. *Journal of Experimental Child Psychology*, *59*(3), 516–548. https://doi.org/10.1006/jecp.1995.1024

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Pylyshyn, Z. W. (1978). Cognition and consciousness in nonhuman species: Open Peer Commentary and Authors' Responses. *The Behavioral and Brain Sciences*, *4*, 592–593.

Rafetseder, E., O'Brien, C., Leahy, B., & Perner, J. (2021). Extended difficulties with counterfactuals persist in reasoning with false beliefs: Evidence for teleology-in-perspective. *Journal of Experimental Child Psychology*, *204*, 105058. https://doi.org/10.1016/j.jecp.2020.105058

Rafetseder, E., & Perner, J. (2012). When the alternative would have been better: Counterfactual reasoning and the emergence of regret. *Cognition & Emotion*, *26*(5), 800–819. https://doi.org/10.1080/02699931.2011.619744

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, *114*(3), 389–404. https://doi.org/10.1016/j.jecp.2012.10.010

Redshaw, J. (2014). Does metarepresentation make human mental time travel unique? *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(5), 519–531. https://doi.org/10.1002/wcs.1308

Redshaw, J., & Suddendorf, T. (2016). Children's and Apes' Preparatory Responses to Two Mutually Exclusive Possibilities. *Current Biology*, *26*(13), 1758–1762. https://doi.org/10.1016/j.cub.2016.04.062

Redshaw, J., & Suddendorf, T. (2020). Temporal Junctures in the Mind. *Trends in Cognitive Sciences*, *24*(1), 52–64. https://doi.org/10.1016/j.tics.2019.10.009

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., & Apperly, I. A. (2006). Children's Sensitivity to Their Own Relative Ignorance: Handling of Possibilities Under Epistemic and Physical Uncertainty. *Child Development*, *77*(6), 1642–1655. https://doi.org/10.1111/j.1467-8624.2006.00964.x

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, *8*(9), 657–661. https://doi.org/10.1038/nrn2213

Schidelko, L. P., Proft, M., & Rakoczy, H. (2022). How do children overcome their pragmatic performance problems in the true belief task? The role of advanced pragmatics and higher-order theory of mind. *PLOS ONE*, 27.

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task. *Frontiers in Psychology*, *12*, 4573. https://doi.org/10.3389/fpsyg.2021.703238

Suddendorf, T. (1999). The rise of the metamind. In *The descent of mind: Psychological perspectives on hominid evolution* (Vol. 25, pp. 218–260). Oxford University Press. https://linkinghub.elsevier.com/retrieve/pii/S0960982214013670

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel and is it unique to humans? *Behavioral and Brain Sciences*, *30*, 299–351.

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers Can Attribute Second-Order Beliefs. *Developmental Psychology*, *30*(3), 395–402. https://doi.org/10.1037/0012-1649.30.3.395

Tulving, E. (1985). How Many Memory Systems Are There? *American Psychologist*.

Tulving, E. (2005). Episodic Memory and Autonoesis: Uniquely Human? In *The missing link in cognition: Origins of self-reflective consciousness* (pp. 3–56). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195161564.003.0001

Uprichard, B., & McCormack, T. (2019). Becoming Kinder: Prosocial Choice and the Development of Interpersonal Regret. *Child Development*, *90*(4), e486–e504. https://doi.org/10.1111/cdev.13029

Van Duijvenvoorde, A. C. K., Huizenga, H. M., & Jansen, B. R. J. (2014). What is and what could have been: Experiencing regret and relief across childhood. *Cognition and Emotion*, *28*(5), 926–935. https://doi.org/10.1080/02699931.2013.861800

Weisberg, D. P., & Beck, S. R. (2010). Children's thinking about their own and others' regret and relief. *Journal of Experimental Child Psychology*, *106*(2–3), 184–191. https://doi.org/10.1016/j.jecp.2010.02.005

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*. *75*(2), 523–541.

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

## Supplementary Material

### Method

### Design

One of three female experimenters (E) tested the children in two separate test sessions. The sessions were at least one day and at most 14 days apart (*M* = 5.39 days apart, *SD* = 3.44). Test session A examined Theory of Mind (ToM) abilities and lasted 10 – 15 minutes while test session B examined Mental Time Travel (MTT) abilities lasting 20 – 25 minutes. The order of session A and B was counterbalanced between subjects. The session format was an interactive online study conducted via a video conferencing platform (primarily *BigBlueButton*, rarely *Zoom* in case of connection issues), where tasks were presented remotely on a computer, laptop, or tablet screen (for details of this general procedure, *see* Schidelko, Schünemann et al., 2021). The tasks were embedded in videos displayed in full screen mode on the child's screen via the conference platform and required the children to answer verbally. The child was able to see herself and E via the webcam in the beginning and in the end of the test session. During the tasks, the child and E were able to communicate through audio streaming only as tasks were presented in full screen mode. In the beginning of the first test session, E apprised the child that she could repeat a question if the child had any comprehension difficulties. At the outset of the first test session, the caretaker verbally gave consent to the child's participation in the study and to the recording of video and audio during the test sessions. This verbal consent was recorded and stored separately from the recording of the two test sessions. Furthermore, the caretaker and child were advised that they can abort the participation at any given moment.

### Participants

The final sample comprised one-hundred-and-twenty 3- to 8-year-old children (36 – 107 months, *M* = 71.51 months; 61 female) consisting of twenty 3-year-olds (aged 36 – 46 months; *M* = 40.85 months, *SD* = 2.70; 10 female), twenty 4-year-olds (aged 48 – 59 months; *M* = 53.95 months, *SD* = 3.65; 10 female), twenty 5-year-olds (aged 60 – 71 months; *M* = 64.95 months, *SD* = 3.38; 10 female), twenty 6-year-olds (aged 72 – 83 months; *M* = 77.90 months, *SD* = 3.54; 11 female), twenty 7-year-olds (aged 85 – 95 months; *M* = 89.75 months, *SD* = 3.40; 9 female), and twenty 8-year-olds (aged 96 – 107 months; *M* = 101.65 months, *SD* = 3.47; 11 female). E1 tested 75 children, E2 tested 23 children and E3 tested 22 children of the final sample. Twenty additional children were tested but excluded from data analyses because of uncooperative behavior ($n$ = 1), language issues ($n$ = 1), procedural issues in the form of distracting siblings or parental interference ($n$ = 2), technical issues during the test sessions ($n$ = 8), concentration difficulties resulting in an error rate at or above 50% in the control questions of the first-order MTT task or in the first-order change-of-location tasks ($n$ =6) or because they did not show up to the second test session ($n$ = 2).

**Material**

Generally, if the child gave a wrong answer to a control question, the question was repeated and if the child again gave a wrong answer, E corrected their response.

**First-order MTT task**

First, the children were introduced to a monkey whom they were supposed to help catch bananas (Figure 1). The bananas fell down from two trees, one with two holes and a wall beneath it (50% chance of catching a banana) and another one with just one hole beneath it (100% chance of catching a banana). First, the children were familiarized with the bananas falling into the holes and then disappearing (3 trials). Afterwards, they watched as the monkey tried to catch a banana on his own (3 trials) by placing the bucket in either one of the three differently colored holes (*see below* for counterbalancing procedure). Then, the children were asked to help the monkey catch a banana. They received two test trials and were asked the following test question: "In which hole should the monkey put the bucket, so that he catches a banana for sure?" Children answered by naming the color of the hole.

**Figure 1.** First-order MTT task material for the first test trial with the question: "In which hole should the monkey put the bucket, so that he catches a banana for sure?".

Counterbalancing Procedure

The first three familiarization (no bucket) trials followed the same order across participants:

- Familiarization trial 1: 100% - red; 50:50 - green | blue

- Familiarization trial 2: 50:50 - red | green; 100% - blue
- Familiarization trial 3: 50:50 - red | blue; 100% - green

The following variables varied in the last three (with bucket) familiarization trials:

- Side of wall: Banana on uncertain (50%) side fell down on the left or the right side of the wall
- Order in which bananas fell from tree: banana on the certain (100%) side fell first or banana on the uncertain (50%) side fell first
- Color of hole in which bucket was placed and successfully caught a banana
- Color of the certain hole in the last familiarization trial: single (100%) hole was red, green or blue
- Success/ no success in last familiarization trial: banana was caught or not caught in last familiarization trial
- If success in last familiarization trial: banana was either caught on certain (100%) side or uncertain (50%) side

The side of the certain (100%) hole in the last familiarization trial was not counterbalanced but always on the left side of the screen. This was done as the certain (100%) hole in the first test trial was always on the right side of the screen, so that the monkey wouldn't look towards the correct choice (i.e. the certain hole) when entering the scene in the first test trial.

Fixed test trials across participants:

- Test trial 1:
  left: 50:50 red | blue; right: 100% - green. After the children chose the hole, they wanted to put their bucket it, the banana on the 50:50 tree always dropped first and into the blue hole
- Test trial 2:
  left: 100% blue; right: 50:50 green | red. After choosing the hole, the banana on the 100% tree dropped first and afterwards the 50:50 banana dropped into the red hole

**Second-order MTT**

The task was based on the pig and discs task (McCormack et al., 2018) and tested if children could reason counterfactually about double determined events. Two test trials with different scenarios were implemented in an animated video accompanied by a narrating voice.

The children first received familiarization trials and answered control questions. First, children watched how a white snowball quickly rolled down a steep hill and then how a grey snowball slowly rolled down a flat hill. They then watched as the two snowballs rolled down the hill at the same time with the white snowball arriving first. Next, a snowman appeared in the valley between the hills and was destroyed by the white snowball rolling down the steep hill (Control Question 1). The event is repeated with the grey snowball rolling down the flat hill and destroying the snowman (Control Question 2). After that, both snowballs rolled down the hills at the same time and the snowman was again destroyed (Control Question 3). Then, a tree appeared at the steep hill, which stopped the white snowball from rolling down the steep hill and thus from destroying the snowman (Control Question 4). In the additive test trial and in the subtractive test trial, again, both snowballs rolled down the hills and the snowman got destroyed (Test Questions, Figure 2). At the corresponding points in the story, the following control and test questions were asked while the video was paused and the still image of the scene was still visible:

- Control Question 1: "What just happened when the white snowball rolled down the hill?" [correct answer: the snowman was destroyed]
- Control Question 2: "What just happened when the grey snowball rolled down the hill?" [correct answer: the snowman was destroyed]
- Control Question 3: "What just happened when both snowballs rolled down the hill?" [correct answer: the snowman was destroyed]
- Control Question 4: "What happens when the tree is standing there on the steep hill and the white snowball rolls down the hill?" [correct answer: the snowman was not destroyed/remained intact]
- Test Question (subtractive scenario): "If the white snowball had not rolled down the hill this time, would the snowman then be broken?" [correct answer: yes]
- Test Question (additive scenario): "If there had been the tree on the steep hill this time, would the snowman then be broken?" [correct answer: yes]

Importantly, the snowman was irreparably destroyed by either one of the snowballs. Thus, in each trial a new snowman was introduced that looked slightly different. Furthermore, the outcome was dichotomous: The snowman was either broken or not.

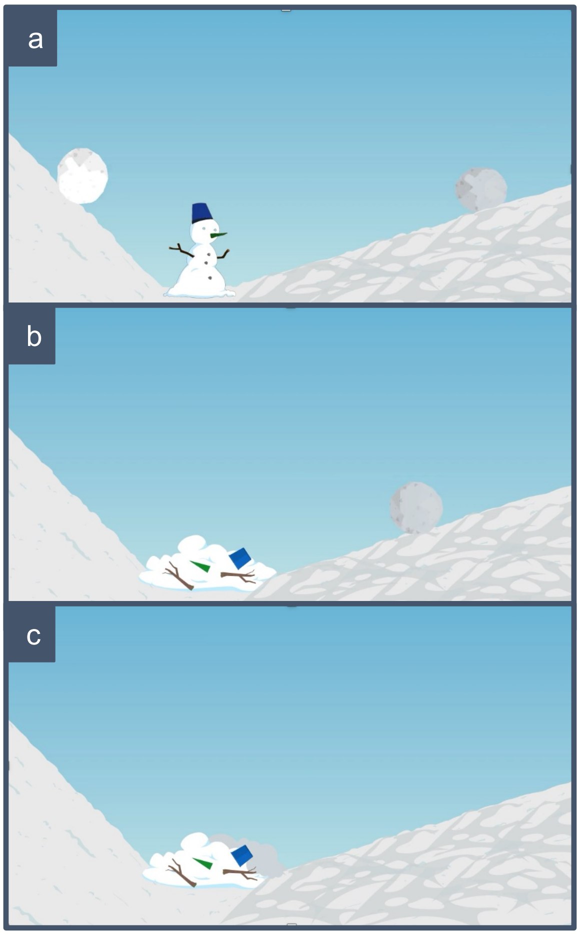



**Figure 2.** Second-order MTT task material for the subtractive scenario. Both snowballs rolled down the hills (a). The quick white snowball destroyed the snowman first (b), then the grey snowball ran into the resulting pile of snow and was destroyed as well (c). Children were then asked the test question: "If the white snowball had not rolled down the hill this time, would the snowman then be broken?". The additive scenario was the same, except that it was a different snowman and an animated finger pointed at the steep hill while the test question was asked: "If there had been the tree on the steep hill this time, would the snowman then be broken?".

**Third-order MTT**

The task was based on the boxes task (McCormack & Feeney, 2015). First, the children were familiarized with a smiley scale that they could use to express their emotions. The base smiley scale included three faces, a sad face, a neutral face, and a happy face (Figure 3a). The base scale could be extended to the left by an even sadder face (Figure 3b) or to the right by an even happier face (Figure 3c). The children learned how to use the extendable smiley scale in two familiarization trials. Children expressed the emotion by naming the background color of the face they were referring to.

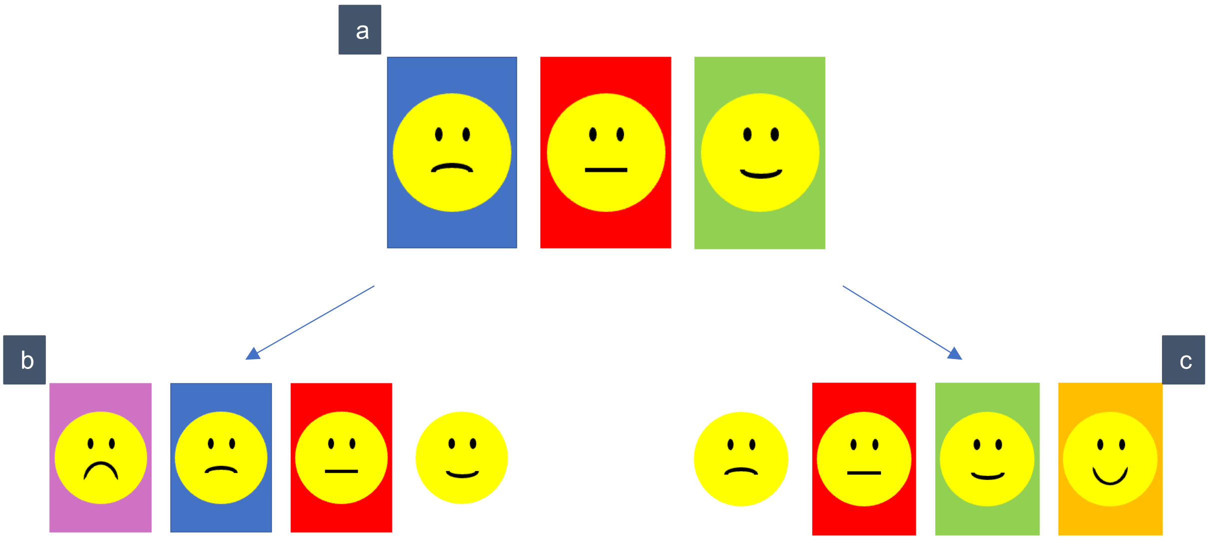



**Figure 3**. Base smiley scale with a blue “sad face”, a red “not happy and not sad face” and a green “happy face” [a]. Extension with a purple “even sadder” face [b]. Extension with an orange “even happier” face [c].

Then the children received one control trial and one experimental trial. In the control trial the children saw two boxes and could choose which one they would like to open. They then received one coin from their chosen box which was added to their collection of coins from the previous games. Then, they were asked: “How do you feel now that there was one coin in your box?”. Children rated their emotion on the base smiley scale by naming the background color of the face they were referring to or by naming the emotion (e.g. “happy”). The emotion they chose became the anchor for the next question: “This is how you felt, when one coin was in your box. Imagine, that there was also one coin in the other box, and you did not win it. How would you now feel about winning this box with one coin? Sadder, happier or the same?” (with ‘this box’ referring to the box they chose and the corresponding faces appearing as the emotions were named, see Figure 4). The other box remained closed, and the children did not learn what was truly inside.

The experiemntal trial was the same as the baseline trial only differing in the colors of the boxes and the test question. The children again first rated their emotion after winning one coin and were then asked: “This is how you felt, when one coin was in your box. Imagine, that there were five coins in the other box, and you did not win them. How would you now feel about winning this box with one coin? Sadder, happier or the same?”.

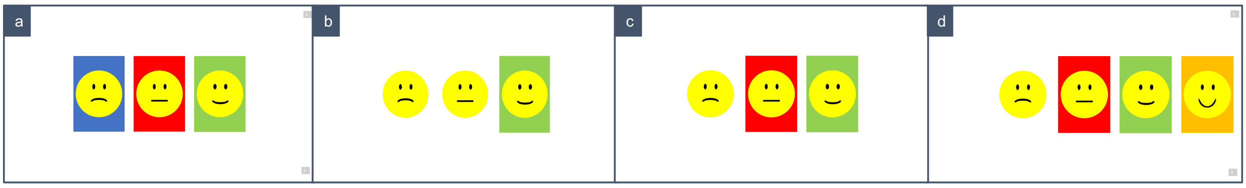


**Figure 4.** Third-order MTT task material for the two baseline and test questions and the corresponding still images on the screen. Example answer: “How do you feel now that there was one coin in your box?” [a]. Child answers: “Happy” [a → b]. Test question: “This is how you felt, when one coin was in your box. Imagine, that there was also one coin in the other box, and you did not win it. How would you now feel about winning this box with one coin?” [b] “Sadder…” [c], “happier…[d]”, “or the same?”.

Coding of Third-Order Mental Time Travel Task

For both trials (control and experimental), children received a score for their initial emotion rating (happy = 1, neutral = 0, sad = -1) and a score for their emotion change rating (sadder than before = -1, same = 0, happier than before = +1). Children’s ratings were summed up for each trial respectively. If this final score was lower in the experimental trial than in the control trial, the child was classified as passer.

Example 1: Child answered the first baseline emotion rating in the control trial with “happy” (1) and the emotion change test question in the control trials with “same” (0). The child, thus receives a baseline score of 1 (1 + 0 = 1). The same child answered the first baseline emotion rating in the experimental trial with “happy” (1) and the emotion change test question in the experimental trial with “sadder” (-1) and thus received a test trial score of 0 (1 + (-1) = 0). Then, the two scores were compared (control trial: 1, experimental trial =0). The child score in the experimental trial was lower than in the control trial. The child was thus classified as passer.

**Theory of Mind Tasks**

Children saw three story lines that tested for their first-order False Belief (FB) understanding (two test questions), their second-order FB understanding (two test questions) and their third-order FB understanding (one test question).

**First Story line (First-order ToM)**

In this version of the standard-change-of-location FB task, the child saw that Protagonist A placed his object O (a ball) in one of two boxes (box 1). Protagonist A then left the scene and Protagonist B entered the scene, relocated O to box 2 and left again. After that, the following control and test questions were asked.

- Control Question 1: "In which box did A place O in the beginning?" [correct answer: box 1]
- Control Question 2: "In which box is O now?" [correct answer: box 2]
- First-order FB Test Question 1: "When A comes back, where will A look for O first?" [correct answer: box 1]

**Second Story line (First-order and second-order ToM)**

In this version of the change-of-location FB task, protagonists A and B watched Object O (a cat) going into box 1. Protagonist A then left the scene. Protagonist B watched as O moved to box 2 and the following control and test questions were asked:

- Control Question 1: "In which box was O in the beginning?" [correct answer: box 1]
- Control Question 2: "In which box is O now?" [correct answer: box 2]
- First-order FB Test Question 2: "When A comes back out of the house, where will A look for O first?" [correct answer: box 1]

The story line then continued as Protagonist B left the scene and Protagonist A re-entered the scene. Protagonist A then learnt that O is in box 2 now. After that, Protagonist B re-entered the scene, the video froze, and the children were asked the following control and test questions:

- Control Question 3: "O is now in box 2. Did A see that?" [correct answer: yes]
- Control Question 4: "Did B see that A saw that?" [correct answer: no]
- Second-order Test Question 1: "Where does B think A will look for O first?" [correct answer: box 1]

**Third Story line (Second-order and third-order ToM)**

Children listened to the video "The school soccer team" (adapted from Liddle & Nettle, 2006) and answered test question after that.

Shortened excerpt from "the school soccer team" (adapted from Liddle & Nettle, 2007)

Max and Paul are best friends. Their favorite thing to do is playing soccer together. Max and Paul both want to play on the school soccer team. But Max thinks that he cannot play soccer as well as Paul. Max thinks that the soccer coach is just choosing Paul for the school soccer team. But the soccer coach thinks that Max and Paul are both good soccer players. He wants* both of them to play on the school soccer team. But the soccer coach knows that Max does not think that he will make the team.

- Second-order FB Test Question 2: Which sentence is true?

  A) Max does not know that the soccer coach wants Max and Paul to both play on the soccer team.* [correct answer]

  B) Max knows that the soccer coach wants Max and Paul to both play on the soccer team.*
- Third-order FB Test Question: Which sentence is true?

  A) The soccer coach thinks that Max believes that he wants that Max to play on the soccer team.*

  B) The soccer coach thinks that Max believes that he wants that Max does not play on the soccer team.* [correct answer]

*German translation with that-complement for want ("möchte, dass")

Children chose a sentence by naming the color of the image on the screen or by repeating the sentence (*see* Figure 5).

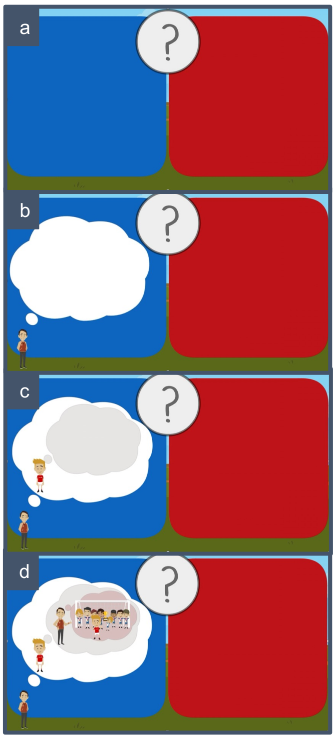



**Figure 5.** Still images of the animations of the third-order ToM test question in the story "the school soccer team". A voice slowly read out the two selectable sentences, which were backed by an associated animation that appeared on the screen (e.g., "The soccer coach thinks [a → b] that Max believes [b → c] that he wants that Max to play on the soccer team [c → d]")

## References

Liddle, B., & Nettle, D. (2006). Higher-Order Theory of Mind and Social Competence in school-age Children. *Journal of Cultural and Evolutionary Psychology*, *4*, 215–229. https://doi.org/10.1556/JCEP.4.2006.3

McCormack, T., & Feeney, A. (2015). The development of the experience and anticipation of regret. *Cognition & Emotion*, *29*(2), 266–280. https://doi.org/10.1080/02699931.2014.912201

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, *114*(3), 389–404. https://doi.org/10.1016/j.jecp.2012.10.010

Schidelko, L. P., Proft, M., & Rakoczy, H. (2022). How do children overcome their pragmatic performance problems in the true belief task? The role of advanced pragmatics and higher-order theory of mind. *PLOS ONE*, 27.

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task. *Frontiers in Psychology*, *12*, 4573. https://doi.org/10.3389/fpsyg.2021.703238

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