THEORY OF MIND DEVELOPMENT IN SWEDISH PRESCHOOLERS

Relations with Language, Executive Function, Temperament, and the Social Environment.

Isac Sehlstedt

Doctoral Dissertation in Psychology
Department of Psychology
University of Gothenburg
March 2024

Cover picture: Taken by Elisabeth Glaas and made into silhouettes by Isac Sehlstedt.
Cover layout: Isac Sehlstedt
Description of the cover picture: The silhouettes are my and Elisabeth's daughters, captured as they hold hands.

Printing: Stema Specialtryck AB, Borås, Sweden, 2024
ISSN: 1101-718X Avhandling/Göteborgs universitet, Psykologiska inst.
ISBN: 978-91-8069-641-8 (PRINT)
ISBN: 978-91-8069-642-5 (PDF)
http://hdl.handle.net/2077/79764

Dedicated to Gösta and Tomas,
you are sorely missed.

Abstract

Sehlstedt, I. (2024). *Theory of Mind Development in Swedish Preschoolers: Relations with Language, Executive Function, Temperament, and the Social Environment*. Department of Psychology, University of Gothenburg, Sweden.

Understanding others in social situations is a cornerstone of any lifespan. A part of social understanding comes from appreciating other's intentions, desires, and knowledge, which can be called an understanding of others' Theory of Mind. However, the measurement of Theory of Mind has predominantly been performed using cross-sectional designs and one type of Theory of Mind test, measuring false belief. Other alternatives that capture a scale measure of Theory of Mind better reflecting a continuum of development across a wider age range are now available. The current thesis investigates this scale longitudinally in relation to previously affirmed, but also less or unexplored, individual and social factors. In brief, the present dissertation finds limited support for individual but some support for social factors. The crucial finding is that Theory of Mind is only marginally related to the investigated factors, apart from Theory of Mind itself. Three studies support the conclusions put forth. *Study I* is a psychometric investigation of the Theory of Mind scale in Swedish preschoolers ages 3–5. The scale was psychometrically examined longitudinally as a 3- and 4-step scale in separate age groups (i.e., at three, four, or five years of age) and for boys and girls, respectively. The results showed that the scale was longitudinally consistent for both versions of the scale. Concerning the separate age groups, the scale was reliable as a 3-step scale in almost all investigated groups. However, the 4-step scale was only reliable when including all age groups (i.e., 3–5-year-old children). This suggests that Theory of Mind scales that include more than three steps might not be appropriate for all preschool ages. *Study II* predominantly investigated the Theory of Mind scale in relation to individual factors, namely executive function, productive language, and temperament. Socioeconomic status was included as a control variable. The individual factors related to Theory of Mind ability were executive function (when analyzed against the 3-step scale) and the temperament variable Shyness (both for the 3- and 4-step scales). Socioeconomic status was also related to ToM at three years of age. *Study III* investigated relations between ToM development and social factors: socioeconomic status, number of siblings, and parental use of mental state words (i.e., mention of cognition, emotion, or desire words). The children's executive function and productive language were included as control variables. Parental use of cognition words was most often found to be related to Theory of Mind, but emotion and desire words were also related, to a lesser extent. In addition, the parents' frequency of spoken cognition words and emotion vocabulary size were related to a faster Theory of Mind development in children. Socioeconomic status and children's productive language were also associated with ToM at four years of age.

In summary, social factors received continued support as factors in Theory of Mind development. However, barely any individual factors surfaced in controlled analyses with Theory of Mind. With a specific focus on longitudinal studies of the development of children's ability to understand other minds, the current thesis uniquely contributes to our understanding of Theory of Mind development in the preschool ages.

Keywords: theory of mind, psychometrics, mental state talk, temperament, executive function, socioeconomic status, productive language

Isac Sehlstedt, Department of Psychology, University of Gothenburg, P.O. Box 500, 405 30 Gothenburg, Sweden. Email: isac.sehlstedt@psy.gu.se

Swedish Summary

När ett barn öppnar sina ögon för att se sina föräldrars ansikten för första gången börjar en social utveckling som kommer fortgå under decennier. Denna utveckling har identifierats som mycket intensiv i förskoleåldern och den tidiga utvecklingen kan påverka ens sociala förmåga långt senare i livet. Därför har forskare ägnat mycket tid och energi åt att söka svar på vad som främjar och motverkar barns sociala utveckling. Den samhälleliga nyttan av denna forskning kan sammanfattas med att nutidens samhällsklimat, och ens eget livs lycka och framgång, till stor del vilar på en god social förståelse. En brist på social kompetens kan därför leda till både privata och professionella problem avhängiga av den aspekten. En fungerande social förmåga är således en essentiell del i att vara människa, och påverkar fler delar av ett liv än den lämnar oberörd.

Ända sedan tidigt 80-tal har många fokuserat på studiet av barns förmåga att förstå andras önskningar, intensioner, och kunskap. Sammantaget har detta beskrivits som att man studerar barnens mentaliseringsförmåga eller Theory of Mind. Trots att forskning pågått i nästintill ett halvt sekel så saknas två aspekter i de allra flesta studier, nämligen upprepade mätningar av mentaliseringsförmåga där man samtidigt har tillgång till upprepade mätningar av erkända (och mindre kända) relaterade förmågor och förutsättningar. Det är alltså sällan man följt samma barn över flera år, samtidigt som man undersökt mentaliseringsförmåga och andra intressanta faktorer. Dessutom har mentaliseringsförmågan ofta mätts med liknande test, ämnade för att mäta en del av mentaliseringsförmåga, nämligen falsk föreställning (eller false belief). Många tidigare mentaliseringsmätningar kan därför ha gått miste om värdefull detaljerad information gällande barnens utveckling. På senare tid har forskare utvecklat test som bättre kan fånga stegvis mentaliseringsutveckling. Bristen på information gällande den individuella utvecklingen av mentaliseringsförmåga och relaterade förmågor kräver en stor forskningsinsats. Vi tog avstamp i ett stegvist test på mentaliseringsförmåga och ämnade att brygga detta informationsgap.

Målet med denna doktorsavhandling är att bättre förstå samband mellan barnens individuella och sociala faktorer och utvecklingen av mentaliseringsförmågan. Tidigare forskning har visat att mentaliseringsförmåga haft samband med några individuella förmågor, som att ha ”många bollar i luften” (m.a.o., exekutiv funktion), språk, och temperament. Relativt många positiva samband mellan exekutiv funktion och språk har rapporterats, men desto färre gällande temperament. Därutöver har sociala förutsättningar och förmågor

som socioekonomisk nivå, familjestorlek (med fokus på syskonskaran), och föräldrarnas förmåga att tala om andra personers tankar, känslor, och begär (m.a.o., mentaliseringsprat) också rapporterats som en del i barnens mentaliseringsutveckling. Tidigare forskning visar starkast positivt samband med socioekonomisk nivå och mentaliseringsprat, medan familjestorlek rapporterats som mindre positivt relaterad till mentaliseringsutveckling. Vår undersökning av mentalisering, individuella, och sociala faktorer delades upp i tre studier med fokus på en av de tre faktorerna.

Det är viktigt att nämna att alla deltagare som är med i de tre studierna har deltagit i samma longitudinella projekt. Därav är det många deltagare som är med i alla tre studier. Överlappet mellan studier är dock inte totalt, då alla studier inkluderar olika många deltagare för varje år. Vi avgränsade också studierna till de familjer som hade svenska som förstaspråk i hemmet. Dessutom genomfördes det longitudinella projektet under åren 2016–2020. Detta innebar att sista årets mätningar avbröts i förtid på grund av Covid-19 pandemin. Därav är deltagarantalet för mätningarna vid fem års ålder betydligt lägre än åren innan.

Studie I var en metodstudie som utvärderade en skala på stegvis mentaliseringsutveckling. Huvudsakliga frågan var om skalan kan pålitligt mäta svenska barns mentaliseringsförmåga. Vi träffade 130 barn som var tre år gamla och mätte deras mentaliseringsförmåga varje år, till och med att de fyllt fem år. Alla barnen deltog inte alla år, utan vid fyra år testades 118 barn, och vid fem års ålder testades 49 barn. Barnens mentalisering uppskattades med hjälp av en vida använd skala som ännu inte utvärderats longitudinellt i Sverige. Skalan kallas, helt enkelt, för mentaliseringsskalan.

När man testar barnen med mentaliseringsskalan får de lyssna på berättelser med tillhörande bilder, dockor, och andra objekt. Barnen får under berättelserna svara på frågor gällande det som hände i berättelserna. Skalan innehöll fyra olika berättelser som krävde att barnen skulle visa att de förstod vad andra kan (1) föredra, (2) tro, (3) veta, och (4) andras falska föreställningar (eller på engelska false belief). Den första berättelsen mätte förmågan att förstå att andra kan föredra saker som man själv inte föredrar. I detta fall, att vissa kan föredra att äta en morot framför kaka. Den andra berättelsen mätte förmågan att förstå att man kan tro olika. I detta fall, att barnet kan tro att en katt har gömt sig i en buske, medan andra kan tro att den har gömt sig i ett garage. Den tredje berättelsen mätte barnens förmåga att förstå att de själva ibland vet vad andra inte vet. I detta fall, att andra inte kan veta vad som finns i en omärkt låda innan de tittat i den. Fjärde berättelsen mätte förmågan att förstå att andra kommer att ta beslut med stöd av

vad de själva vet, och att de inte alltid kan veta vad barnet vet. I detta fall var det en berättelse där barnet fick se att det i en plåsterförpackning låg, istället för plåster, en spik. Barnet skulle då gissa vad andra skulle tro det fanns i förpackningen (m.a.o., ett test som mäter barnets förmåga att förstå falsk föreställning eller false belief).

Våra resultat visade att mentaliseringsskalan med fyra berättelser (m.a.o., berättelserna om vad andra föredrar, tror, vet, och falska föreställningar) var pålitlig för svenska barn i 3–5 års ålder. Dock visade det sig att mentaliseringsskalan var mer instabil när vi analyserade enskilda åldrar. Därför utvärderade vi om en kortare skala som endast inkluderade de första tre berättelserna (m.a.o., berättelserna om vad andra föredrar, tror, och vet) kunde vara mer lämplig i yngre åldrar. Det vi fann var att mentaliseringsskalan med tre berättelser var stabil vid både fyra och fem års ålder.

Slutsatsen av våra resultat från *Studie I* blev således att mentaliseringsskalan fungerar väl för att mäta svenska förskolebarns mentaliseringsförmåga. Det är dock viktigt att noggrant överväga hur många berättelser som är lämpliga för de åldersgrupper som man avser att undersöka. Efter att ha bekräftat att vi kan lita på att vår valda mentalieringsskala kan fånga utvecklingen hos svenska barn, ville vi undersöka vilka förmågor som har samband med mentaliseringsförmågan barnet uppvisar.

Studie II undersökte de individuella förmågorna exekutiv funktion, språk (aktivt ordförråd), och temperament hos barnet. Vi inkluderade även det sociala måttet på socioekonomisk nivå som en kontrollvariabel. Exekutiv funktion mättes med ett sorteringstest där barnen skulle sortera kort baserat på antingen färg eller form. Det svåra med uppgiften var att när man sorterade enligt färg, så sorterade men inte efter form, och vice versa. Barnets språkbruk och temperament bedömdes av föräldrarna med hjälp av standardiserade formulär. I denna studie inkluderade vi 121 deltagare vid två års ålder, 121 deltagare vid tre års ålder, och 111 deltagare vid fyra års ålder. Det vi fann var att exekutiv funktion vid två års ålder och socioekonomisk nivå hade positivt samband med mentaliseringsförmåga vid tre års ålder. Dessutom hade blyghet vid två års ålder ett negativt samband med mentaliseringsförmågan två år senare.

Exekutiv funktion och språkförmåga har i tidigare forskning varit de faktorerna med starkast samband med mentaliseringsförmåga. Trots det anser vi att blyghetsfyndet är minst lika starkt enligt *Studie II*. Blyghetsfyndet var nämligen det enda fynd som visade samband med mentalisering när vi kontrollerade för tidigare mentalisering (m.a.o., mentalisering vid tre års ålder).

Med de individuella faktorerna granskade fokuserade vi på att undersöka sociala faktorer i samband med mentaliseringsförmåga.

Studie III undersökte barnens socioekonomiska nivå, deras familjestorlek, och deras föräldrars mentaliseringsprat eller mental state talk. För att kunna mäta föräldrarnas mentaliseringsprat lät vi föräldrarna och barnen sitta ensamma i ett rum tillsammans och prata om bilder i en bilderbok. Bilderna hade mer eller mindre tydliga tankemässiga, känslomässiga, eller behovsstyrda/begärliga budskap. Föräldrarnas mentaliseringsprat delades upp i tre dimensioner: (1, frekvens) hur ofta föräldrarna använde mentaliseringsord, (2, proportioner) hur många mentaliseringsord de använde i relation till det totala antalet ord och (3, vokabulärstorlek) hur många olika mentaliseringsord de använde. Anledningen till att vi delade upp föräldrarnas mentaliseringsprat i tre olika mått är att frekvenser har använts mest tidigare, men proportioner har fördelen att kompensera för hur länge eller snabbt föräldrarna talar med barnen. Vokabulärmåttet användes för att utvärdera om även detta, som tidigare inte undersökts, har något samband med barnens mentaliseringsförmåga (då det tidigare visats ha samband med förståelse för andras känslor). I denna studie inkluderade vi 82 deltagare vid tre års ålder, 82 deltagare vid fyra års ålder, och 33 deltagare vid fem års ålder.

Våra resultat från *Studie III* visade att alla typer av mått på föräldrarnas mentaliseringsprat (m.a.o., frekvens, proportion, och vokabulär storlek) hade samband med barnens mentaliseringsförmåga. Gällande frekvens var det föräldrarnas förmåga att tala om andras tankemässiga reflektioner när barnen var två år som hade ett positivt samband med hur snabbt barnens mentaliseringsförmågan utvecklades. Även vid granskning av vokabulärmåtten var det storleken på föräldrarnas känslomässiga vokabulär när barnen var tre år som hade samband med hur snabbt barnens mentaliseringsförmåga utvecklades. Det är viktigt att nämna att båda dessa fynd syntes när vi kontrollerade för tidigare mentaliseringsförmåga, att de var statistiskt tydliga, men små i faktiskt uppmätta värden.

Några fynd som syntes i analyser där tidigare mentaliseringsförmåga inte inkluderades i analysen bör också nämnas. Resultat gällande proportioner av föräldrarnas mentaliseringsprat visade att prat om tankemässiga reflexioner hade negativt samband vid två års ålder med barnens mentaliseringsförmåga vid fyra års ålder. Däremot, tankemässiga reflektioner hade positivt samband vid tre års ålder med barnens mentaliseringsförmåga vid fyra års ålder. Dessutom hade föräldraprat om behovs-/begärrelaterade reflektioner vid tre års ålder negativt

samband med mentaliseringsförmåga vid fyra års ålder. Därutöver hade också socioekonomisk nivå samband med mentaliseringsförmåga vid fyra års ålder i frekvens och proportionsanalyserna, och barnens språkbruk vid två års ålder var relaterat till mentaliseringsförmåga oavsett vilken analys som genomfördes (m.a.o., frekvens, proportion, eller vokabulär storlek).

Det negativa sambandet mellan prat om behov/begär och mentaliseringsförmåga har rapporterats tidigare. Man tror att det negativa sambandet kan förklaras med att föräldrar som fokuserar på att prata om begär med äldre barn gör att barnen inte får lika mycket erfarenhet av de svårare perspektiven där man ska förstå vad andra kan tänka och tycka. Det oväntade var proportionsfyndet gällande prat om tankemässiga reflektioner, som visade negativt samband vid två års ålder med mentalisering vid fyra års ålder. Dock vill vi förklara det på liknande sätt som gäller för behov-/begärresultaten. Det kan vara så att föräldrar kan hjälpa barnen förstå andras perspektiv genom att ofta prata om vad andra kan tänka eller tycka. Likväl kan det vara viktigt att ge nog med kontext med ord som inte är mentaliseringsord. Speciellt före tre års ålder. Det kan givetvis även vara så att barnens egen förmåga att ta sig an en social situation eller social information om andras tankar kan påverka vad föräldern pratar om. Tyvärr har vi inte kunnat utvärdera vilken av dessa förklaringar som är mest gångbara i denna doktorsavhandling. Det vore dock intressant att undersöka i framtida studier.

Sammantagningsvis gav studierna ett svagt stöd för att individuella faktorer var av större vikt för mentaliseringsförmåga än de sociala. Istället tyder våra resultat på att utvecklingen av mentaliseringsförmåga är erfarenhetsbaserad, med ett fokus på sociala erfarenheter. Därav kan vi tolka barns mentaliseringsförmåga som en förmåga som utvecklas i samband med den sociala miljö som barnet finner sig i. Framtida studier uppmuntras att undersöka hur barn påverkar deras sociala miljö och hur det i sin tur påverkar barnens sociala förmåga.

Preface

This thesis is based on the following three papers, which are referred to by their Roman numerals:

I. Sehlstedt, I., & Hjelmquist, E. (2024). *Theory of mind development in Swedish preschoolers: A longitudinal investigation.* Manuscript.

II. Sehlstedt, I., & Hjelmquist, E. (2024). *Developing Theory of Mind in Relation to Executive Function, Socioeconomic Status, Language and Temperament.* Manuscript.

III. Sehlstedt, I., Hansson, I., & Hjelmquist, E. (2024). *The longitudinal relation between mental state talk and theory of mind.* Manuscript under review.

The studies in this thesis were financially supported by a grant from the Swedish Research Council (VR, grant nr. 2014-18190-113123-31).

Acknowledgements

It is necessary and a pleasure to salute all who contributed to my journey and this finished thesis.

Firstly, I am forever grateful to all the families that participated. Thank you all for your time, effort, and outstanding devotion.

Secondly, I would like to thank:

Tomas Tjus, my late main supervisor (1954-2018), for all you did and wanted to contribute. You made the project engaging, and enjoyable. I am forever grateful.

Sara Landström, for briefly stepping in as my main supervisor after Tomas's death, for your disproportionally large contribution to this thesis. I am more grateful to you than you'll ever know.

Marek Meristo, my current main supervisor. It was an unforgiving task to come in late to a finished project. Still, I am thankful for everything you contributed to the thesis.

Erland Hjelmquist, my perpetual co-supervisor, for steadfast supervision throughout the years and for the countless insightful conversations and discussions. Your contribution has been invaluable.

Stefan Hansen, my previous examinator and now co-supervisor, for repeatedly helping me find perspective on the thesis and constantly making me better at presenting research in general.

Linda Hassing, my examiner, for listening to all my minor and major questions and for your support in all the work that followed with your answers.

Kerstin Falkman Watson, for going above and beyond what could ever be asked of you and making a world of difference for me as an aspiring researcher and person.

Leif Strömvall, Jesper Lundgren, Ann Backlund, Anne Ingeborg Berg, and Martin Lövdén, for your thoughtful assistance and guidance in my times of need.

Mikael Heimann, Jakob Åsberg, Jan-Eric Gustavsson, and Maria Gröndahl, for essential input and comments on the thesis.

Rochelle Ackerley, Ulf Dahlstrand, Valgeir Thorvaldsson, Daniel Bergh, Magnus Lindwall, Kajsa Hansen Jang, Pernilla Larsman, Stefan Franzen, Marcus Praetorius Björk, Isabelle Hansson, for enjoyable and fruitful discussions regarding statistics and proper scientific procedure.

Therese Wallstedt, and Tommy Reinholdsson, for ample assistance with many aspects of the project.

ICON-Lab, AMBLE, SOCEF, ADA-Gero, MEDTEC WEST, Sahlgrenska/Berget members, for widening my scientific perspective.

Oscar Hagsten, Tobias Thyi, Adrian Cederqvist, and Olle Bergenfeldt Thorén, for devotedly exploring the project data and giving me new insights.

Ioanna Blasko and Helen Hjort, for transcribing many of the parent-child conversations.

Karin Strid, for repeatedly lending me a plethora of items I lacked.

IT support, for indispensable assistance throughout the project.

I would also like to thank:

David Norlin, David Neequaye, Caroline Järdmo, Malin Joleby, Sofia Calderon, Hanna Larsson, Johan Skoog, Joel Gerafi, Carl-Christian Trönnberg, Jonas Burén, Fanny Gyberg, Lina Nyström, Ann-Sofie Sten, Jonas Gillenstrand, Jeremy Ray, David Sandberg, Sandra Buratti, Emelie Ernberg, Timothy Luke, Manuela Ravazdi, Angelica Hagsand, Marja Önsjö, Ida Malm, Stefan Winblad, Toms Voits, Marcus Lindqvist, Erik Nilsson, Gaia Olivo, Bodil Karlsson, Lina Wirehag Nordh, Petra Boström, Dönmez Aziz-Kaan, Pär Stern, Petra Löfgren, Marie Eckerström, Amos Pagin, Ask, Andrea Karlsson Valik, Therese Björkäng, Anna Baadsgaard, and Sara Sjödin, for many relaxing, thoughtful, and joyous conversations. Thank you for the memories. May your efforts always see favorable outcomes.

Erik Mac Giolla, Olof Wrede, Lukas Jonsson, Py Eriksson, Emma Ejelöv, Magnus Bergquist, Patrik Michaelsen, Gustaf Glavå, and Simon Skau, my heart and my mind are substantially enriched thanks to you recurrently inspiring me to join a hefty gym session, a talkative walk, a peaceful yoga class, a scenic bike ride, a table tennis rally, or to run after a football enthusiastically. I cannot thank either of you enough.

Thirdly, thank you to my friends outside the walls of the workplace for your interest, understanding, and constant support of my work. You contributed more to the current thesis than you might believe.

Och givetvis, min familj. Tack till:

Min pappa Gösta (1944–2021), för ditt beständiga och villkorslösa stöd. Jag tänker på dig varje dag.

Min mamma Eva, mina syskon Petra och Frank, svärföräldrarna Marie och Stigbjörn, svägerskorna Jenny och Cecilia, och svågrarna Staffan och Henrik, för allt, för alltid.

Min livspartner Elisabeth, för ditt stöd inför alla delmål som tagit oss hela vägen hit, för att du är lugnet i alla stormar, och för all den kärlek du ger mig. Jag älskar dig.

Slutligen, tack Elma och Kristin, för alla pussar, kramar, och fina samtal. Pappa älskar er.

Isac Sehlstedt
Gothenburg, March 2024

Table of Contents

List of Abbreviations

Abbreviation	Definition
CFB	Content False Belief
CFI	Comparative Fit Index
DB	Diverse Belief
DCCS	Dimensional Change Card Sort task
DD	Diverse Desire
EF	Executive Function
EFB	Explicit False Belief
FB	False Belief
FBU	False Belief Understanding
FIML	Full Information Maximum Likelihood
HE	Hidden Emotion
KA	Knowledge Access
LGCM	Latent Growth Curve Models
MLR	Maximum Likelihood Estimator
MST	Mental State Talk
RMSEA	Root Mean Squared Error of Approximation
SES	Socioeconomic Status
SRMR	Standardized Root Mean Square Residual
ST	Simulation Theory
TLI	Tucker Lewis Index
ToM	Theory of Mind
TT	Theory Theory

List of Tables

List of Figures

Introduction

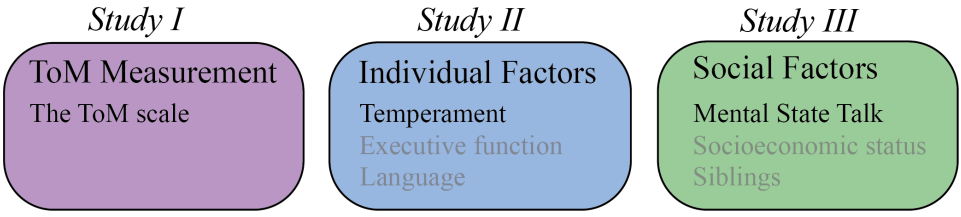
The current thesis aims to deepen our knowledge about how social cognition development in young children relates to individual and social factors. For this purpose, Theory of Mind (ToM) development was investigated in Swedish preschoolers with a language level allowing verbally-based testing and communication. According to Röska-Hardy (2009):

> … [ToM] denotes the conceptual system that underlies the ability to understand, predict and interpret the thoughts, feelings, and behavior of self and others by reference to specific mental states (states of mind). […][it] is used to refer to (1) the ability to impute mental states, i.e., to mentalizing or mind-reading…, (2) the study of children's understanding of the mind in developmental and cognitive psychology, and (3) the "Theory Theory" account of mental state attribution. (p. 4064).

All three parts of the ToM definition above are incorporated in the current thesis. The first is by measuring ToM using a test that requires the participant to attribute mental states to be successful. The second is by holistically incorporating the study of ToM development, where many related factors are investigated simultaneously. And the third is a part used to guide the thesis theoretically.

However, the studies in the thesis are not separated to address the stepwise definition presented above. Instead, the investigations into ToM development were divided into three studies with three aims (see Figure 1, visualizing the conceptual framework and focused variables per included study): the measurement of ToM, individual factors related to ToM, and social factors associated with ToM. More specifically, the first study investigates the ToM scale developed by Wellman and Liu (2004), a well-known instrument used to assess ToM development, and here aimed to ensure its applicability in a Swedish context. The second study mainly investigates ToM from a longitudinal perspective and how individual factors such as executive function (EF), language, and a recently introduced variable in ToM research, temperament, may contribute to ToM development. Finally, the third study investigates the predictive power of predominantly social factors, focusing on parents' mental state talk (MST) on their children's ToM development. The following introductory text provides summaries of previous research on these topics.

Figure 1 - *The Conceptual Framework Guiding this Thesis.*



Note. ToM = Theory of Mind; The Figure presents the focal variables in each included study, with temperament being the central factor in *Study III*, and Mental state talk being the main factor in *Study III*. An additional social factor (i.e., Socioeconomic status) is included in *Study II*, and additional individual factors (i.e., productive language and executive function) are included in *Study III* as controls. Notably, the grayscale of a variable signifies the relative focus of the included variables in the study, with gray variables being less focused than black variables.

Theory of Mind

Premack and Woodruff (1978) originally defined the term ToM and stated that "An individual has a theory of mind if he imputes mental states to himself and others" (p. 515). In their study, Premack and Woodruff claimed that Sarah, a chimpanzee, could understand others' perspectives. However, according to Dennett (1978), Sarah was trained to seem like she had a ToM. Dennett (1978), therefore, outlined philosophical and theoretical arguments to guide ToM research, thus giving us the early main focus of ToM research, namely studies on false belief (FB; Baron-Cohen et al., 1985; Wimmer & Perner, 1983).

To understand that someone has a FB, the child needs to grasp that another person can believe something is correct, even though the child knows it is incorrect (Ruffman, 2014). For instance, a common way to test FB is to present a child with an object (e.g., a box of band-aids) and reveal that the box contains something unexpected (e.g., a plastic coin). Then, introduce another character (e.g., Dolly) and ask the child what Dolly will find in the box. If the child can say that Dolly believes there are band-aids in the box, then the child understands FB. Testing of FB has become widely used for investigating ToM-ability partly due to its possibility for varying the complexity of tasks (Ruffman, 2014; Wellman et al., 2001). For instance, the FB test can measure the ability to understand what a second person might know about the beliefs of a third person, and so on.

Measures of ToM

There have been a few recent efforts to summarize the methods used to evaluate ToM development during preschool (Beaudoin et al., 2020), later childhood (Osterhaus & Bosacki, 2022), and over the two first decades of life (Fu et al., 2023). For instance, Beaudoin et al. (2020) summarized more than thirty years of research and reported considerable heterogeneity but a few key commonalities. They exemplified that ToM development measures in the preschool age can capture the development in several sub-domains (e.g., emotion understanding, desire understanding, belief understanding, or knowledge). Although measuring ToM within a sub-domain might be better than measuring one type of test (e.g., the relatively advanced Faux pas test; Happé, 1994), only measuring within the sub-domains still limits the scope of the investigation into the development of ToM understanding (Beaudoin et al., 2020; Wellman & Liu, 2004).

Some researchers have suggested an alternative to capture ToM development across sub-domains (Carlson et al., 2013; Hiller et al., 2014; Wellman, 2002). Therefore, there are now batteries or scales designed for this purpose (Beaudoin et al., 2020), namely comprehensive measures using questionnaires (e.g., the Theory of Mind inventory by Hutchins et al., 2012), or comprehensive measures using direct measures (e.g., the ToM scale by Wellman & Liu, 2004).

A Scale for ToM

Beaudoin et al. (2020) report that one of the most evaluated and validated comprehensive measures using direct measures is the ToM scale (Wellman & Liu, 2004), which includes (amongst other items) tests of beliefs, desires, knowledge, and FB. The ToM scale has been used to investigate ToM development globally with children from two years of age (Hiller et al., 2014), up to late teens (Wellman et al., 2011a). The scale has a maximum of eight steps but it is most common to use five steps with children younger than six years of age (Pava, 2019). Children at this age typically learn to understand the inner lives of others in a specific order (Etel & Yagmurlu, 2015; Sundqvist et al., 2018; Wellman & Liu, 2004). Around three years of age, (1) a child can grasp that others might have unique preferences (or diverse desires; DD), (2) and personal beliefs that differ from their own (diverse beliefs; DB). At around 3.5 years of age, (3) a child can understand that there are situations where other individuals are missing important information that allows them to make correct decisions (knowledge access; KA). Between the ages of three and six, (4) a child can gather that a person may believe incorrect information is correct and subsequently act upon that incorrect information, resulting in suboptimal behavior (content false belief; CFB; Wellman et al., 2001; Wellman & Liu, 2004). Between the ages of four and seven, (5) a child starts to understand that a person might be showing an emotion that is incongruent with the emotion the person is experiencing (hidden emotion; HE).

The Outline of the Thesis at Hand

Specific measures were applied initially by Wellman and Liu (2004) that can be used to evaluate the ToM scale. The measures aim to analyze the stability and reliability of the difference in difficulty between each step included in the scale. The measures are often excluded in previous ToM scale studies, and studies with the measures often need to exclude steps of the scale to find acceptable reliability and stability. Moreover, earlier scale steps are appropriate for 2–3-year-old

children to avoid ceiling effects. In contrast, the later steps (e.g., the more challenging steps than HE) are appropriate when children are five years of age or older to avoid floor effects. *Study I* includes a more nuanced analysis of the reliability and stability of the scale than has been done previously.

Examples of factors that have been proposed to be closely tied to ToM development are language development and EF. With regards to language, there is a wealth of research describing the associations between language and ToM ability (de Villiers, 2007, 2021; Miller, 2006; Milligan et al., 2007; Ruffman, 2014; Ruffman et al., 2003). Likewise, EF is often found to be related to ToM development (Carlson et al., 2002, 2013; Wade et al., 2018). Some researchers even indicate that EF is a prerequisite for ToM, while others suggest EF is more remotely related to ToM (for a review, see Moses & Tahiroglu, 2010). Therefore, the relations between these individual factors and ToM will be explored in *Study II*.

Temperament is a less well-studied individual factor that might affect ToM development. Previous research has suggested that some temperament dimensions (e.g., being shy or active) are related to ToM (Lane & Bowman, 2021). However, the relations do not have consistent support across studies, and only a few studies have investigated the relationship between the ToM scale and temperament (Mink et al., 2014; Song et al., 2016). In addition, only a few longitudinal studies have investigated FB in relation to temperament (Brink et al., 2015; Carlson et al., 2004; Selcuk et al., 2018; Suway et al., 2012; Wellman et al., 2011b). Only one longitudinal investigation of the ToM scale has been performed with children 2–5 years of age; in that study, aggression and ToM were measured over roughly a year (Song et al., 2016). There is, therefore, a lack of longitudinal investigations into the age-related relations between the ToM scale and broader measures of temperament in the preschool years. Thus, *Study II* will further these investigations.

With the individual factors reviewed, social factors are related to ToM ability. Social factors relevant to the current thesis are socioeconomic status (SES, often measured as parental/maternal educational level), number of siblings, and the types of words parents use when conversing with their children (Devine & Hughes, 2018). Parental education has been suggested to be most important to ToM development, provided that one includes a wide range of socioeconomic levels. Still, a measure of education is relevant in studies with less variation. Family size or the number of siblings is another measure pertinent to the development of ToM (Devine & Hughes, 2018; Perner et al., 1994) where having

a sibling, and sometimes having several siblings, is reported relevant for ToM development (Hou et al., 2022; Prime et al., 2016, 2017). Finally, the mental state talk (or MST) parents use when conversing with their children have been suggested to be linked to ToM development (Tompkins et al., 2018). For instance, asking questions and conversing about what others might believe or prefer is commonly found to be related to later ToM ability. These social factors are important and interesting to investigate together with ToM development. This topic, with a focus on MST, is covered in *Study III*.

Repeated measurements of the same children as they grow older are preferable for development studies. A longitudinal study of ToM may be a study that includes a measure of a cognitive ability that is interesting for the development of ToM at an early age, and ToM at a later age (e.g., Mink et al., 2014). But the power of a longitudinal investigation into ToM comes from repeated measures (i.e. more than one measurement) of both ToM and other abilities of interest. Surprisingly, there is only one longitudinal study in which ToM was measured more than twice (Wellman et al., 2011a). Additionally, many factors that have been investigated in relation to the ToM scale have mainly been investigated at one point in time and not repeatedly. A longitudinal investigation into the relations between a child's ToM scale performance and the same child's performance on other tests before and after repeated measurements of the ToM scale might give new insights.

In sum, the current thesis will longitudinally investigate ToM development using the ToM scale in relation to language, EF, temperament, and social factors. The first section of the introduction will outline some of the most prominent theoretical positions in current ToM research. Later sections will further discuss how previous research efforts have related ToM development to language development, EF, temperament, and social factors.

Theories of Theory of Mind

The theoretical landscape related to ToM is vast, and the overlap between theories is sometimes considerable. Premack and Woodruff (1978), the founders of ToM as a concept, presented their theoretical standpoint by stating that ToM is ”[…] to our knowledge, universal in human adults. Although it is reasonable to assume that their occurrence depends on some form of experience, that form is not immediately apparent. Evidently, it is not that of an explicit pedagogy. Inferences about another individual are not taught, as are reading and arithmetic; their acquisition is more reminiscent of walking or speech.” p. 525. Their account has also inspired many theories presented after that.

The Combination of Simulation and Theory - Theory

For the current thesis, two approaches stand out: *Simulation Theory* (ST; Harris, 1992, 2009) and *Theory Theory* (TT; Gopnik & Wellman, 1992; Wellman, 2014). The critical difference between the two is that ST suggests a tangible learning process, and TT suggests a more abstract learning process (Apperly, 2008; Miller, 2016; Röska-Hardy, 2009; Tanaka, 2017). As will become apparent, there is no perfect example of how to tease these theories apart (Apperly, 2008). Still, an effort will be made below.

ST can be summarized as “You first have to know yourself to know others.” It suggests that we use our own experiences to understand others. In other words, we figuratively put ourselves in the other’s shoes to understand their minds (Harris, 1992). Based on this swap, we automatically and intuitively infer what is happening in the other person’s mind. The child’s experiences will limit the simulation they can make. A child can only simulate what they have experienced (Röska-Hardy, 2009). In sum, as the child ages, they will likely have more diverse information when setting up the simulation.

The point of departure for TT, by contrast, is the general and overarching theories of the human psyche that we form to understand the minds of others. These evolving theories gain sophistication as children grow older. In other words, TT suggests we infer the current mindset of the target person by applying abstract theories based on our past experiences (Apperly, 2008). The target person’s response is determined by methodically utilizing knowledge about the person and how general mental processes function (Gopnik & Wellman, 2012; Wellman, 2014). In short, TT states that hypotheses about a person are carefully created and tested, enabling learning.

Designing studies that unequivocally support ST or TT has not been easy. Therefore, researchers have suggested that the two theories may complement each other (Apperly, 2008; Harris, 1992; Hughes & Dunn, 1998; Mitchell et al., 2009). Mitchell et al. (2009) explicitly suggest that ST is more applicable in the early stages of ToM development, while TT develops later. ST contributes to understanding others' beliefs, desires, and knowledge, while TT better explains FB. According to this hypothesis, FB tests are best handled by keeping reasoning more theory-driven (like TT) and less simulation-focused (like ST). It might be easier to refrain from overtly reporting one's perspective if the reasoning about the FB scenario is less tied to one's perspective.

Gopnik and Wellman (2012) argued that TT lacked a strong computational foundation and suggested that Bayesian statistics might be highly relevant for TT. They stated that Bayesian statistics and TT involve expecting outcomes and revising models or theories in light of unexpected experiences or results. Wellman (2014) further reviews research and states that the computational aspects of Bayesian modeling strongly support TT (Goodman et al., 2006). He suggests that children make predictions about others, correct predictions in light of unexpected results, and develop better ToM. Wellman (2014) uses this reasoning to separate ST from TT, as ST does not center around theory building or refinement. However, Wellman (2014) could not present any findings specific to the ToM scale but rather findings related to learning (including social learning). The reasoning by Wellman (2014) might be valid for general social learning, but the question is if it holds specifically for ToM development.

Since Gopnik and Wellman's proposal to view social development through the lens of Bayesian modeling, three studies have investigated ToM development from this viewpoint (Asakura & Inui, 2016; Baker et al., 2016; Wang et al., 2019). None of these three studies support TT. However, Asakura and Inui (2016) investigate the relationship between ST and TT. They published a Bayesian model of ToM (based on the framework used by Goodman et al., 2006), which permits simulation ability (from ST) in the theory component of TT. Asakura and Inui found that their model explained performance on the ToM scale reported in previous research. More specifically, they found that the performance on the early steps of the scale (i.e., DD and KA) distinctly predicted CFB task performance. Moreover, their model was similarly successful when comparing differences in age, countries, or developmental delays. Therefore, the "hybrid" solution (ST/TT) accounts well for ToM scale performance (Asakura & Inui, 2016). This finding does not fit the standpoint initially made for the separation

between the theories by Gopnik and Wellman (1992), opposes the later Bayesian elaboration by Wellman (2014), and fits better with the proposition made by Mitchell et al. (2009), Apperly (2008) and Harris (1992) of compatibility between ST and TT.

In summary, the discussion regarding ST *or* TT, instead of ST *and* TT, has made researchers skeptical of a hybrid solution. It might be necessary to clarify that the way a person thinks, or the way the brain operates can be Bayesian; however, analyses performed using Bayesian statistics do not necessarily support that the brain works in a Bayesian manner. Nonetheless, Asakura and Inui's (2016) model is relevant to how we believe the brain develops ToM. They created a so-called Bayesian causal net model (advocated as an important learning model supporting TT according to Wellman, 2014) on the performance of ToM scale steps. Notably, their model was based on TT, but it included the simulation ability present in ST. Crucially, Asakura and Inui's (2016) model had a high accuracy in explaining performance at ages 3–6 in six different countries and for four types of developmental delays on the ToM scale in previous studies. To be clear, Asakura and Inui did not contrast a ST and a TT model, and they did not compare the performance of the hybrid model to a pure TT model.

Nonetheless, orthodox TT and ST proponents cannot explain the results and conclusions of Asakura and Inui. Therefore, the current thesis will use the hybrid ST/TT framework as an overarching framework without specification in Bayesian terms. In other words, like mastering chess, children start by improving their ability to simulate the behavior, thereby gradually acquiring the ability to create systematic theories about other's behaviors. Alternative theories will be presented in the discussion and compared to the ST/TT account.

Alternative Approaches

Some alternative approaches are essential to mention. For instance, *expression* or the *emergence* of ToM (for a review, see Moses & Tahiroglu, 2010). These approaches focus on the role of EF in the development of ToM, with EF being the accelerator (as in the expression approach) or gatekeeper (as in the emergence approach) of ToM. There is also the *nativist-modular* account (Leslie et al., 2005; Scholl & Leslie, 1999, 2001) where ToM ability is thought to be innate, environmentally cued to develop, and that the limits of ToM development are the same worldwide (Saxe, 2006). Another alternative is Heyes and Frith's (2014) *implicit/explicit* account, where two systems manage ToM development. The implicit system develops autonomously and independently of EF, while the

explicit system develops deliberately and depends on EF. All these theories have merit. However, the frameworks they provide do not fit the current thesis as well as the ST/TT hybrid. Specifically, the ST/TT hybrid is better positioned to capture a broader range of possible findings regarding the development of ToM in the preschool years. Nonetheless, all mentioned frameworks will be compared in the discussion, focusing on each framework's explanatory value to the current thesis's findings.

With the ST/TT hybrid as a background, an introduction of concepts and functions related to ToM in previous research should be presented. Therefore, the sections leading up to the method section will be structured as follows. First, a description of relevant findings and discussions related to the concept will be presented. Its relation to ToM will be discussed, and finally, the measure used to capture that concept will be introduced.

Language Development

Tremendous developments occur in the first year of life with regard to language (Kuhl, 2004, 2011). The first month can be without signs of development, but the child still absorbs the rules of the languages spoken at home (Hernandez et al., 2000; Werker & Tees, 1999). Furthermore, infants need months to understand and maybe a year to say their first word (Kuhl, 2004). Systematically and statistically (Kuhl, 2004), the child learns to capture the properties relevant to the language (or languages) spoken around them (Hoff, 2006). Importantly, as summarized in Hoff (2006), hearing a language spoken around you is not enough for proficient language development. Instead, input directed to the child encompassing all levels of language, i.e., speech(sound), structure, and meaning, of any of the world's languages, and the opportunity to use that language in interaction with others is required (Kuhl, 2004). Deficiencies at these levels can negatively affect language development (Hoff, 2006; Kuhl, 2004)[1].

The Fundamentals of Language

Aspects of language relevant to the text below are grammar, comprehension, and production. Grammar is a very general concept and encompasses at least rules for word formation, morphology in general, and syntax (rules for how to construct phrases, clauses, and sentences, Teleman et al., 1999). Comprehension refers to all aspects of the reception of spoken language in the early stages of development, often focusing on understanding words. Production refers to the child's active use of spoken language, likewise often referring to the production of words. It seems as if productive vocabulary and grammar development begins with a long one-word period, and before 12 months of age, it is hard to find children with a productive vocabulary larger than approximately ten words (Bates et al., 1995; Eriksson & Berglund, 1999). Around the age of 16 months, word production increases rapidly, enabling grammatical development (Bates et al., 1995; Eriksson & Berglund, 1999). This rapid increase continues until the child's second birthday (Borgström et al., 2015).

Additionally, Berglund and Eriksson (2000) showed that productive vocabulary and grammar proficiency (i.e., "…morphological markers for the possessive form, definite singular, definite plural, plural marking, and past tense or supine"., p. 135) were highly correlated, especially around two years of age.

[1] Language can develop without auditive input using only the visual modality (as with sign language).

Therefore, the relationship between vocabulary and grammar can be considered close, with little to no dissociation (Bates et al., 1995; Berglund & Eriksson, 2000). However, it is necessary to mention major differences between individuals regarding the speed of vocabulary acquisition. For example, Berglund and Eriksson (2000) showed that 2-year-old children could have a productive vocabulary of less than 100 and others with a productive vocabulary of around 700 words (with a scale measuring 710 words common Swedish words). Notably, that variation in development is mirrored in the variation of grammar skills, with some that have not developed any measurable grammar skills and others close to proficient in all the measured aspects of the grammar skills defined above (Berglund & Eriksson, 2000).

Language and ToM

Efficient communication is a significant factor in understanding FB. For instance, children with specific language impairments lag substantially behind their peers' ToM development (Nilsson & de López, 2016). Perhaps the most striking example comes from studies with deaf children (e.g., Schick et al., 2007; Siriattakul et al., 2021; Wellman et al., 2011a). Research on these children shows that deaf children born into households with hearing parents have a slower ToM development than hearing children in hearing households (for a summary, see Siriattakul et al., 2021). However, deaf children born into households with deaf parents show very similar ToM development to hearing children in households with hearing parents (Meristo et al., 2007). For the current thesis, participants who could participate in oral communication without learning or hearing difficulties were included.

Concerning research investigating associations between ToM and orally conveyed language, Milligan et al. (2007) have published the most comprehensive meta-analysis available. They found that all language factors measured were significantly related to FB with medium (i.e., .34–.66) mean effect sizes. General language measures (e.g., an experimenter in person measuring various production, comprehension, and syntax using a series of structured questions) were found to be significantly better at predicting ToM in comparison to receptive vocabulary (e.g., a [Peabody] Picture Vocabulary Test). However, general language measures, semantics (e.g., synonym judgment task), and syntax (e.g., the complexity of grammar items in forms) were comparable predictors of FB. Milligan et al. (2007) further reported metanalytical results showing that effects held for many language abilities and FB tests, suggesting

generality in the relations between language and FB. The findings by Milligan et al. (2007) were supported in a more recent meta-analysis suggesting that general language (i.e., a combination of production, comprehension, and syntax) might be the most appropriate language measurement to use with young children (Farrar et al., 2017). However, Farrar et al.'s (2017) meta-analysis only analyzed language measures against Explicit False Belief (EFB) performance, which is a ToM scale step that is commonly excluded from Guttman analyses of the ToM scale based on analyses performed and recommendations by the original authors (Wellman & Liu, 2004).

Even though there are no summarized studies investigating associations between ToM and productive vocabulary, some studies on the topic are worth bringing to light. Durrleman et al. (2022) performed a vocabulary intervention and found no relation between FB and vocabulary training. Longobardi et al. (2021) cross-sectionally investigated children's ability to name objects and actions performed on pictures and correlated productive language ability with ToM performance. They found that productive language and ToM were related. Other longitudinal studies have described the relationship further by reporting significant relations between earlier measures of productive language (as measured by questionnaires) and later ToM (Brooks & Meltzoff, 2015; Farrar & Maag, 2002; Watson et al., 2001).

The Language Measure

In search of an appropriate measure of a child's vocabulary, the MacArthur Communicative Development Inventories (MCDI; Bates et al., 1994; Fenson et al., 1994; Marchman & Bates, 1994) was chosen, as it is considered a valid measure of productive language development (Camaioni et al., 1991; Dale et al., 1989) and is available in Swedish (Berglund & Eriksson, 2000; Eriksson & Berglund, 1999). The MCDI is divided into several scales, all with a different age group in focus. It measures the variability in the productive language of a child based on parental (or similar) ratings. Importantly, MCDI does not measure phonology, frequency of utterances made by children, if the child was imitating or spoke spontaneously, or in which/how many different contexts a child has produced a word (Bates et al., 1994). Therefore, keeping reasoning in line with these limitations will improve the quality of the conclusions drawn from MCDI data (Bates et al., 1995).

Measures from the Swedish version of the MCDI when children were two and three years of age are included in the current thesis.

Executive Function

EF is a cognitive ability that allows us to guide our mental processes top-down, thus enabling goal-directed behavior (Espy, 2004; Miller & Cohen, 2001). Consequently, a well-developed EF is paramount to having an efficient and enjoyable life experience (Moriguchi et al., 2016). This claim is supported by EF's relations to various aspects of life, such as mental and physical health, school and job success, and social and relational prosperity (Diamond, 2013).

EF allows an individual to resist being trapped in automatic attention, or so-called bottom-up processes, and to filter incoming information better. More specifically, EF can be described in relation to a scale of attentional processes, namely Alerting, Orienting, and EF (Petersen & Posner, 2012; Posner & Petersen, 1990). Alerting is the attentional process of producing and maintaining high sensitivity to an ongoing task or situation. Orienting is the ability to prioritize information available in the current space based on location or sensory modality. Finally, EF, on the attentional process spectrum, is the ability to manage conflicts regarding information, including emotions, cognitions, and behavior. Additionally, EF itself is commonly divided into three separate but cooperating processes.

The Three Executive Processes

EF is commonly divided into three separate albeit interacting cognitive processes: inhibition, working memory, and cognitive flexibility (Diamond, 2013; Miyake et al., 2000). Inhibition is the ability to ignore irrelevant information, thoughts, or emotions to stay true to one's goals (Diamond, 2013). A lack of inhibitory control results in behavior being more heavily guided by automatic attention-grabbing stimuli or old habits. Therefore, inhibition is central to being efficient, flexible, and coherent. The relation between early inhibitory control and outcomes in later life has been investigated. One study with 1000 participants has shown that children 3–11 years of age with the ability to not be impulsive in everyday situations had better outcomes in adulthood compared to those who did not (Moffitt et al., 2011).

Working memory (WM) is the ability to manipulate and remember information that is no longer possible to perceive (Baddeley, 2000; Baddeley, 1992). Therefore, a typical functioning WM is paramount in all sequential behaviors and tasks (e.g., reading, cooking, math, FB, etc.). Also, the ability to reason or make decisions by weighing the advantages and disadvantages of

different factors would be impossible without WM. WM is qualitatively different from short-term memory (STM). STM is passive storage where information can reside before being sent to long-term memory. STM cannot manipulate information, while WM does (Aben et al., 2012).

Cognitive flexibility (CF) is shifting perspectives or priorities. CF greatly relies on both inhibition and WM to function. Inhibition allows previous perspectives or priorities to be suppressed. In contrast, WM capacity enables the individual to hold online what the previous strategy was and the current course of action. Most tasks designed to capture CF are tasks where two or more rules must be followed. As a result, CF has been found to develop later than the other two cognitive processes included in our EF (Davidson et al., 2006; Garon et al., 2008).

Executive Function and ToM

Research on the relations between EF and ToM has focused on early childhood (Devine & Hughes, 2014; Weimer et al., 2021). The reason for this could be the relative focus on the FB task and when a child is finally able to pass the FB task. As meta-analytically summarized (across 102 studies and 9994 participants) by Devine and Hughes (2014), several factors should be considered when investigating EF and ToM. They reported that (1) EF was strongly related to FB (i.e., mean weighted r was .38) for 3–5-year-old preschoolers, (2) the relation between EF and FB is comparable across many geographic regions, (3) the most common FB tasks relying on the content (e.g., unexpected content of a toy-car in a band-aid box) or location (e.g., that a ball has been moved to another container) were equally associated with EF, (4) that all EF tests (included in the analysis) were associated with FB, (5) associations between EF and FB remained when controlling for verbal ability and age, and (6) composite scores of at least two tests measuring EF and two tests measuring FB revealed a more robust association compared to when either was measured with a single test. The only negative association between EF and ToM that Devine and Hughes (2014) reported was that (7) larger sample sizes were related to smaller effect sizes. In summary, points 1–5 suggest that the relation between EF and FB is very stable. Points 6 and 7 give a perspective that should be considered when investigating EF and ToM relations.

The question that remains is, why is EF related to ToM? One suggestion is that the relation between ToM and EF might be indirect, and improvements in EF may not be linked with improvements in ToM; however, better EF might lead

to an improvement in the quality of the social interaction and increase the possibility of socializing with others (Hughes, 1998; Moses & Tahiroglu, 2010). These differences in opportunity and quality of social interactions, brought about by differences in EF ability, might be what assists ToM development and not EF ability itself (Moses & Tahiroglu, 2010). A more obvious reason would be that handling and comparing two minds, my own and others, implies taxing EF.

The Measure of Executive Function

The measure of choice for EF (or CF) for the current thesis was one of the most widely applied versions of a child-focused EF test (Devine & Hughes, 2014), the Dimensional Change Card Sort task (DCCS; Zelazo, 2006). The test is built up of three stages. During each stage, the children are asked to sort cards into trays. The rule for sorting is clearly stated to the child at each stage, and the challenge is to change the sorting strategy according to a new rule flexibly. The cards to be sorted depict two combinations of color and shape, while the trays have pictures that do not match the color and shape of the cards that the child is supposed to sort. A child can complete the different steps of the DCCS at certain ages. For instance, children can complete the first stage of the DCCS at three years of age but not the second. Children four and a half years old can complete the first two stages but may struggle with the third. Children do not systematically complete the third and last stage until they have reached the age of seven or nine (Davidson et al., 2006). Children struggle to complete the second stage at young ages due to "attentional inertia" (Anderson, 1979; Kirkham et al., 2003). The classical definition of attentional inertia applied to a DCCS task means that the rule to sort according to color gets carried over to the second stage, where all cards will still be sorted according to color instead of shape. This finding of attentional inertia suggests that CF is not developed enough for the inhibition of a previous perspective to be successful (Chatham et al., 2012; Kirkham et al., 2003).

The current thesis includes measures of a Swedish version of DCCS when children were two, three, and four years of age.

Temperament

A general definition of temperament that integrated the research-based insights accumulated since an earlier definition by Goldsmith et al. (1987) was suggested by Shiner et al. (2012), specifically "Temperament traits are early emerging basic dispositions in the domains of activity, affectivity, attention, and self-regulation, and these dispositions are the product of complex interactions among genetic, biological, and environmental factors across time" (p. 437). It could be noted, though, that what dispositions or dimensions of behavior should be included in a framework for temperament has varied widely over the years. Still, four of the more prominent theories will be mentioned below.

Prominent Theories of Temperament

The first theory might be one of the earliest theories of temperament. Thomas, Chess, et al. (1960) identified nine dimensions of temperament: activity level, approach-withdrawal, threshold of responsiveness, persistence or attention span, adaptability, distractibility, quality of mood, intensity of reaction, and rhythmicity. Even though some of the dimensions defined by Thomas, Chess, et al. (1960) do still carry some clinical relevance (Shiner et al., 2014), most dimensions have been found to have low internal consistency, were difficult to discriminate from each other conceptually, and the suggestion was to reduce the number of dimensions to describe temperament better (Roberts & DelVecchio, 2000; Sanson et al., 2002). The second theory, Goldsmith's theory (Goldsmith et al., 1987), can be criticized for including many dimensions (as it includes as many dimensions as Thomas & Chess's account). Additionally, Goldsmith's theory is mostly applied to infancy. However, Goldsmith's theory for children past infancy has been developed into Rothbart's theory (e.g., Goldsmith & Rothbart, 1991). The third theory, Rothbart's theory (Rothbart & Bates, 2007; Sanson & Rothbart, 1995), includes three broader dimensions, namely: Reactivity, or *Negative affectivity* (e.g., negative mood, irritability, anger), Self-regulation, or *Effortful control* (e.g., non-distractibility, or persistence), and Approach-Withdrawal, Sociability, or *Surgency* (e.g., approach to novel situations). The fourth theory, Buss and Plomin's (1975; 1984) influential theory of early temperament dispositions, originally included three dimensions, namely: *Emotionality* (e.g., displaying emotion), *Activity* (e.g., active approach to activities), *Sociability* (e.g., preferring social games). A fourth dimension, *Shyness* (e.g., taking a long time to warm up to people), was later added (Buss &

Plomin 1984). A fifth dimension, *Impulsivity* (e.g., often switching between tasks), was later added, then removed due to a lack of empirical evidence for a genetic relation. However, impulsivity now has empirical support (Gagne & Saudino, 2010), making it fit Buss and Plomin's theory.

Temperament and Theory of Mind

How children approach and handle familiar and new social situations sets the stage for their own experience (Lane & Bowman, 2021). In other words, a child's propensity to dive into a social situation will result in a relatively large amount of experience of other minds, and a lack of interest in social interaction will set hard limits on the ability to get even indirect experiences of social interactions and other minds. Therefore, individual differences in temperament are a possible explanatory factor for understanding children's ToM ability. Some of the previously reported associations between temperament and ToM are described below.

Shyness and False Belief

It has been suggested that such basic dispositions as temperament may influence ToM development (Lane & Bowman, 2021). No meta-analysis has summarized the relation between temperament and FB, and previous research presents a varying pattern. For instance, no significant relationship between temperament measures and FB was reported by Calero et al., 2013, Carlson et al., 2004, and Colonnesi et al., 2010). However, the social or shyness dimensions have been one of the most frequent significant predictors of FB, sometimes together with other temperament dimensions (for a review, see Lane & Bowman, 2021). For example, LaBounty et al. (2017) found that shyness was positively related to FB (effect size was large, e.g., $\beta s = 0.48$). However, Walker (2005) also reported that lower shyness or withdrawn behavior scores were related to higher FB scores, but only for boys. To complicate the positive relation between shyness and FB, Walker also reported that girls exhibiting high prosocial behavior were related to high FB scores. Wellman et al. (2011b) reported regression analyses revealing that non-aggressive, shy/withdrawn, and perceptually sensitive temperament at three years of age was related to higher FB scores at five years of age, even when controlling for IQ, inhibition, gender, and FB at three years of age. Noteworthy, none of Wellman et al.'s (2011b) zero-order correlations (when calculated using their sample size and correlation coefficients) were significant (i.e., $p<0.05$). Lane et al. (2013) reported that high social withdrawal (together

with low cortisol levels) was related to high ToM. In other words (without implying causality), children who were socially withdrawn but remained calm and relaxed in that socially withdrawn situation had higher FB scores.

Other Temperament Dimensions and False Belief

There are also indications that other temperament dimensions might be related to ToM development, namely inhibition and activity. For instance, Longobardi et al. (2017) reported a significant positive relationship between inhibition to novelty and ToM in 4–5 year-olds but no significant relationships in a group of 3–4 year-olds. Moreover, higher activity is related to lower FB (LaBounty et al., 2017).

Temperament and the ToM Scale

Studies with only FB measures aside, some relations between temperament and the ToM scale have been reported previously. For instance, Mink et al. (2014) found that shyness was predictively (from 18 months) and concurrently positively related to ToM at three years of age (effect size was moderate, $\beta s = 0.31$). Korucu et al. (2017) showed that effortful control (that includes measures of inhibitory control) was positively related to ToM scale scores in a large cross-sectional study with 3–6-year-old children. However, effortful control was the only dimension of temperament they included, not all three dimensions that are parts of Rothbart's theory. Concerning inhibition, only one previous study seems to have been performed. Suway et al. (2012) found that high behavioral inhibition and negative peer interaction at two years of age were each predictive of low ToM at three years of age.

Regarding activity, Mink et al. (2014) reported that activity level at 18 months was negatively related to ToM scale scores at three years of age (effect size was moderate, $\beta s = -0.34$); however, the relationship between activity level and ToM was heavily influenced by outliers, making it irrelevant (Mink et al., 2014). Nonetheless, Henning et al. (2011) found cross-sectional support for a negative relation between activity and ToM scale score for 3–to 6-year-old children. In sum, no clear conclusion can be reached from previous studies, but there are indications that ToM might be related to shyness (or social aspects of temperament), inhibition, and activity.

The Lack of Longitudinal Investigations

The bulk of the previous studies investigating the relationship between temperament and ToM have been cross-sectional, and only a few have been longitudinal (i.e., Brink et al., 2015; Carlson et al., 2004; Mink et al., 2014; Suway et al., 2012; Wellman et al., 2011b). Notably, few previous studies have investigated the longitudinal relations between temperament and the ToM scale. Other approaches have included habituation time and socially observant behavior in relation to implicit FB (Brink et al., 2015), Rothbart's theory in relation to a battery of FB tasks (Wellman et al., 2011b), a test of intentions, desires, and perspective taking using Goodman's and Rothbart's theory (Carlson et al., 2004). However, Mink et al. (2014) did investigate the relationship between the first three steps of the ToM scale and Rothbart's theory. Suway et al. (2012) also included three original ToM scale steps (with one extra task included). They related them to behavioral inhibition (but not any of the four prominent theories outlined above). Given the limited number of longitudinal studies, more longitudinal research on the relation between the ToM scale and temperament is warranted.

The Temperament Measure

The model by Buss and Plomin (Buss & Plomin, 1975, 1984) includes fewer dimensions that are easier to separate than the larger models of Shiner et al. (2012). Therefore, the chosen temperament measure is the EASI, or EAS Temperament Survey (Buss & Plomin, 1984; Swedish translation Hagekull & Bohlin, 1990). The EASI comprises five subscales: Emotionality, Activity, Sociability, Shyness, and Impulsivity. The questionnaire measures a child's temperament by asking guardians/parents or teachers to rate the child's temperament using five items for each subscale. The first four subscales are reliable, consistent, and stable (Bould et al., 2013; Mathiesen & Tambs, 1999; Walker et al., 2017), but including impulsivity in the scale seemed less appropriate (Walker et al., 2017).

The current thesis includes measures with EASI when the children were two and four years of age.

Social Factors

The remaining factors studied in this thesis are social. For instance, positive child development relies heavily on parent-child interactions (Fay-Stammbach et al., 2014; Madigan et al., 2013; Zimmer-Gembeck et al., 2017). The classical opinion of parenting used to be that parents exerted a unidirectional influence from the parent (and mainly the mother) to the child (Kuczynski et al., 1997). However, children are now regarded active in and a competent part of their development (Kuczynski et al., 1997), and parent-child interaction is now considered a balanced interplay between the parent and the child. Additionally, fathers are now being recognized as contributing uniquely (Marsiglio et al., 2000) and in synergy with the mother (McHale & Rasmussen, 1998) to the child's development (e.g., EF development as reported in Ribner et al., 2022)[2]. The current thesis's three social factors of interest are SES, family size (or the number of siblings), and parental MST.

Socioeconomic Status

SES is a measure intended to capture differences in access to material and social resources, and a combination of factors measures it (e.g., income, occupation, and education) or any of the factors on their own (Buckingham et al., 2014; Hoff et al., 2002). Compound variables of SES are relatively rare, and maternal education has been one of the most frequently used non-compound SES variables (e.g., Ensminger & Fothergill, 2003).

Socioeconomic Status and Theory of Mind

Social factors related to ToM have been summarized from various perspectives in recent years (e.g., Devine & Hughes, 2018; Miller, 2016; Szpak & Białecka-Pikul, 2019; Tompkins et al., 2018), resulting in key insights to many relevant factors. For instance, on average, children in homes with higher SES have been found to perform slightly better on FB tasks (Devine & Hughes, 2018). Summarizing almost 50 studies, Devine & Hughes (2018) reported that the effect was modest but significant. Notably, the effect of SES was significant, albeit attenuated when controlling for verbal ability (Devine & Hughes, 2018).

[2] Importantly, the amount of same sex marriages have increased steadily during the last decades (Kolk & Andersson, 2020) and a wealth of research has shown that children of lesbian mothers or gay fathers show typical development and adjustment (for reviews, see Biblarz & Stacey, 2010; Golombok, 2017; Manning, Fettro, & Lamidi, 2014; Tasker, 2005).

Additional insights provided by Devine and Hughes meta-analysis were that the strength of the association between FB and SES was higher if the study used a compound SES measure (rather than a single measure), if the children were closer to 74 months (i.e., older than six years of age) than 36 months (i.e., three years) of age, and if the study included a wider range compared to a narrower range of ages. They also found that early publications reported stronger correlations between FB and SES than later studies.

The measure of parental SES level was the mean of parental educational attainment ranked on a 7-point scale utilizing the Hollingshead index (Hollingshead, 1975). Education is often used as an SES indicator, also in ToM research. The way of scaling education differs widely, from relying on steps from very basic education to university level, or simply by counting number of years of education, or dichotomizing between high and low education (Devine et al., 2016; Ensor et al., 2014; Jenkins et al., 2003; Meins et al., 2013; Taumoepeau & Ruffman, 2008). The Hollingshead index was chosen as it provides a reasonable differentiation of educational levels and a framework to capture educational attainment reliably and in a systematic and replicable manner. Not least, it is easy for parents to provide the information. The point scale used to capture educational attainment for each parent was divided into (1) Less than nine years primary education, (2) nine years of primary education, (3) high school (or Gymnasium in Sweden), (4) post high school education or (Advanced Higher Vocational Education, Higher Vocational Education or Folk High School in Sweden), (5) Bachelor's degree, (6) Master's degree, and (7) graduate professional training.

The current thesis includes a measure of parental SES when children were two years of age.

Siblings and Social Abilities

Social understanding often develops in interaction with siblings, and that experience may generalize to other relationships (for reviews, see McHale et al., 2012; Parke, 2004; Teti, 2002). For instance, a large-scale study (N = 20649) investigating the ability to negotiate peer relationships (as measured by teacher ratings) in relation to family size revealed that social competence (or ability to keep friends) was, on average, lower at preschool for single children in comparison to children with at least one or two siblings (Downey & Condron, 2004). Furthermore, a follow-up study including 11820 fifth-grade (out of the original 20649 preschool) participants revealed the same pattern and that the

differences had increased in fifth grade (Downey et al., 2015). This means siblings might have an important role in developing social skills.

Siblings and Theory of Mind

Devine and Hughes (2018) found metanalytical support that children with more siblings have a more developed FB (see also Cassidy et al., 2005; Perner et al., 1994). However, they also report that the strength of the association was modest. Nonetheless, the association between the number of siblings (or family size) and FB remained even when controlling for verbal ability when analyzing cross-sectional data or previous FB when investigating longitudinal data. Devine and Hughes's (2018) FB results align with the finding that the presence of siblings is associated with better social understanding and keeping friends when measured by teachers ratings (Downey & Condron, 2004; Downey et al., 2015).

The current thesis includes a measure of the number of siblings when children were two years of age.

Mental State Talk and How it is Measured

The social interaction measure of interest for the current thesis is MST. In simple terms, MST involves using words relating to cognitions (e.g., believe, think, know), emotions (e.g., happy, sad, angry), or desires (e.g., want, like).

Oftentimes, a sentence uttered by a parent may include words included in more than one of the MST categories. Additionally, parents vary in the way they incorporate MST in conversations. Therefore, when measuring MST, there are at least two measures of MST to consider. Absolute frequency, when each time a child hears an MST word is counted (Ruffman et al., 2002; Symons et al., 2006; Van Bergen & Salmon, 2010), or proportions that have the benefit of controlling for the amount of words a parent utters (Howard et al., 2008; Meins et al., 2003). One current issue when evaluating the suitability of proportions or absolute frequency as a measure of MST is that the number of uttered words is not often reported (Tompkins et al., 2018). Therefore, the current thesis evaluates both measures of MST in the same dataset.

Mental State Talk and Theory of Mind

The amount of MST used by parents has been shown in metanalyses to be related to children's performance on ToM tasks (Devine & Hughes, 2018; Tompkins et al., 2018). Devine and Hughes (2018) meta-analyzed data from 28 studies and

reported that the effect between parental MST and FB was modest but significant. The effect was comparable when controlling for verbal ability using a subsample of 12 studies. They also found, analyzing results from six longitudinal studies, that MST was still significantly and moderately related to FB when controlling for earlier FB. Devine and Hughes (and Tompkins et al., 2018) also highlighted that the setting where MST was measured (i.e., unstructured play, looking at pictures, or from a questionnaire) did not influence the relation (but Tompkins et al. did find that a reminiscing session, talking about memories, that was not included in Devine and Hughes's analysis, was significantly less related to FB). Additionally, the relation to FB was lower if the amount of MST was controlled for verbosity (i.e., proportions of MST). Crucially, the studies that report frequency and proportions give mixed results, with some finding relationships to ToM (FB; Moeller & Schick, 2006), and others not (Adrián et al., 2007; Martin & Green, 2005; Symons et al., 2006). However, frequencies and proportions are not the only available MST measures.

One measure that might be overlooked in previous research is the parents' mental vocabulary size. One previous investigation into vocabulary size focused on emotional vocabulary and understanding emotions (Martin & Green, 2005). The spontaneous active use of different MST words may differ between parents. A parent with a broader, more nuanced MST vocabulary may add quality to the MST that further aids ToM development. It seems as if MST vocabulary size is an uninvestigated part of MST research, that could reveal hidden factors related to ToM development.

The current thesis includes measures of parental MST when children were two and three years of age.

Aim of the Thesis

The thesis consists of three empirical studies based on data from the longitudinal project Brain, Mind and Culture: Pathways to Mentalizing, Language, and Reading, planned and performed by the interdisciplinary research group Arena for Mind, Brain, Learning, and Environment (AMBLE). This doctoral thesis investigates ToM development and its relation to other factors that can be expected to be important for understanding the nature of ToM. It consists of three empirical studies investigating three different aspects of ToM development.

The first main research question was: *How can ToM development be reliably measured longitudinally in a Swedish context with a specific ToM scale?* Explicitly, the psychometric properties and reliability of the ToM scale by Wellman and Liu (2004) was investigated using common reliability measures. This research question is addressed in *Study I*.

The second main research question was: *How do individual factors relate to ToM ability in preschoolers?* The aim was to investigate the same children over time to reliably capture development and elucidate how productive language, EF, and temperament relate to ToM. This research question is predominantly addressed in *Study II*.

The last main research question was: *What social factors are related to Theory of Mind development?* The ambition was to study how family size (or number of siblings), SES, and MST relate to ToM ability. This research question is predominantly addressed in *Study III*.

Methods and Materials

The Project

The three studies included in this thesis were based on data from the four measurements performed in the project Brain, Mind and Culture: Pathways to Mentalizing, Language and Reading. The complete project involved investigating the development of ToM in preschool-aged children. Age-adequate tests covering ToM, EF, memory, phonemic awareness, and productive language were measured repeatedly and often with various instruments and occasionally even methods. For instance, some cognitive tests were complemented with concurrent electroencephalographic (EEG) measures or eye-tracking registrations. Tests excluded from the current thesis are Baby Stroop (Hughes & Ensor, 2005), an episodic memory test (Meltzoff, 1985), DUVAN (Wolff, 2013), "the Farmhouse" (based on the Missing Scan task; Buschke, 1963), number repetition from NEPSY (Korkman et al., 1998), Peabody Picture Vocabulary Test (Dunn & Dunn, 2007, 1981), the Serial reaction task (Koch et al., 2020), an EEG task investigating neural responses to language-related auditive stimulation (Leppänen et al., 2011), and the ToM eye tracking task (similar to Surian & Geraci, 2012). Reasons for the exclusions, some preliminary results from the excluded tasks, and a complete list of the tests included in the project are described in Appendix I.

Additionally, two tasks were excluded due to unexpected issues. The Spin the Pots test (or the Lazy Suzan task; Hughes & Ensor, 2005), a measure of EF, was excluded due to seemingly unreliable scores in a large portion of the sample across three years of measurement. Additionally, the Child Behavior Questionnaire – Very short form (Rothbart & Bates, 2007; Sanson & Rothbart, 1995), which includes measures of the temperament dimensions Negative affectivity, Surgency, and Effortful control, was excluded due to it being unreliable in the current sample. A more detailed description of these two tasks and the reason for exclusion can be found in Appendix II.

The current thesis aimed to investigate cognitive development for children aged 2 to 5 using both on-site tasks and off-site forms. A complete list of the tests in the current thesis is summarized at the end of the Descriptive Results -section.

Recruitment

All children were recruited via the Swedish registry, "the Swedish State Personal Address Register" (SPAR), which includes all persons registered as residents in Sweden. The sample included families living in or around the city of Gothenburg (West Sweden) with children born in October, November, and December of 2014 or January or February of 2015. Only the child's age and postal code were used to restrict the sample. The participants received an invitation letter with extensive information about the project and an informed consent form to be signed by both parents. They asked them to provide their email and telephone number. An envelope was provided to ease the return of the informed consent. All who sent in a signed informed consent were contacted by telephone to book a first meeting.

The Sample

Invitations were sent to 2920 parents, almost exclusively mothers, with children meeting the age criterion. A total of 230 families gave informed consent. This means that 7.8 % replied to the invitation. The aim was to test the participants around two years of age. Unfortunately, due to technical issues regarding the testing facilities at the department, assessment could not start until the children were around two years and four months. This four-month lag was also kept at the three other data collection points.

After the first round of measurements at two years of age had been performed (from late December 2016 to the beginning of July 2017), 180 children had been tested. Testing was performed during the same months of the year in all subsequent years. At three years of age, 149 (83%) families participated in testing, and at four years of age, 136 (76%) families participated. At five years of age, on-site testing was suspended in March 2020 due to the Covid-19 pandemic, leading to only 54 (30%) participants being tested out of 130 (72%) who were still interested in participating. Therefore, retention rates were relatively high for all years except the last (83%, 91%, 40%).

When looking at the sample demographics, there are some differences between those who stayed in the study and those who left. There was a tendency for parents with less than a bachelor's degree to leave the project. Also, a family might have been more likely to leave the study if the family had more than one child.

Notably, all participants who could not be tested on-site were sent off-site forms in case they still wanted to participate in some way. A total of 76 forms were sent to those not tested on-site, and 45 were completed.

The sample cannot be considered representative of the population in Sweden. The invitations were sent to families geographically close to the on-site testing facilities to make travel time reasonable for the participants. Therefore, a selection bias towards the West coast (i.e., Västra Götalands län) of Sweden is evident. Additionally, the current sample had a high education level on average. The population average in Sweden is 2% with PhD degrees and 29–42% with a bachelor's degree or higher (*Statistics Sweden*, 2020). Also, the current sample had 7 (4%) parents with PhD degrees, and 61.1% of the total 180 participants had bachelor's degrees or higher. Noteworthy, the proportion of parents with a bachelor's degree or higher rose to 66.5% for the third (and last complete) data collection time.

The On-Site Collection

On-site testing of all participants was performed by the author of this thesis. All testing sessions were planned to take around 90 minutes, including breaks. During testing, parents were routinely asked if they thought the child needed to take a break or abort testing completely. In 42% of the sessions, testing took longer than 90 minutes, but this was often because of many, sometimes lengthy breaks (e.g., snack breaks) during testing. On-site testing time at two years of age took a mean of 98 (SD=21) minutes; at three years of age, 87 (14) minutes; at four years of age, 96 (16) minutes; and at five years of age, 51 (5) minutes. The length of testing time of the first three years made it impractical to allow more than four bookable time slots each workday. Even though participants were allowed to book all days of the week for over six months when they were around two years of age, only 180 could participate.

On-Site Tests

Executive Function Test: Dimensional Change Card Sort task (DCCS)

The Dimensional Change Card Sort task (DCCS; Zelazo, 2006) was used to measure EF. Cards have two dimensions: shape and color. The cards have one out of two shapes (e.g., a rabbit or a boat), and these shapes have different colors (e.g., blue or red). There are only two versions of the cards (e.g., one with a red rabbit and the other with a blue boat). The task has three stages: pre-switch, post-switch, and border stage. During each stage, children are asked to sort the cards based on the rules conveyed by the experimenter. The pre-switch stage is used as a baseline and is carried out as follows. Two shallow trays with pictures (e.g., one tray with a picture of a blue rabbit and one with a picture of a red boat) are placed in front of the child.

The child gets a brief introduction to "A card game," where the child gets familiar with the materials. A practice round is then performed where the experimenter confirms that the child can sort cards in accordance with one dimension (e.g., color). This is done by observing that the child when prompted, puts a card with a red rabbit in the tray with the red boat picture and puts a blue boat in the tray with a blue rabbit picture. If the child can complete the practice round, the experimenter can move on to the test round of the pre-switch stage. In the test round, the participant is asked to sort six cards (e.g., three blue boats and three red rabbits). Each card is presented one at a time. The child is prompted verbally with the relevant dimension by the experimenter whilst being presented with one card and is then asked to place it in one of the trays (e.g., "This card is red, where does that one go?"). If the child sorts five cards (or more) correctly and sorts all six cards, then the child will proceed to the post-switch stage. The post-switch stage starts with an explanatory part, where the experimenter is explicit about the change that is now happening in the game (e.g., "We are now going to play a new game. We are not playing the color game anymore. We are going to play the shape game."). The rules of the Post-switch stage are then explained verbally (e.g., Rabbits go here, and boats go there. So, if you get a rabbit, you place it in that box. And if you get a boat, you put it in that box"). However, the child is not allowed to train on the new rule (as done in the pre-switch stage), and the test round of the post-switch phase starts immediately. As in the pre-switch stage, the participant is asked to sort six cards (e.g., three blue

boats and three red rabbits). Each card is presented one at a time. The child is prompted with the relevant dimension verbally by the experimenter while being presented with one card and is then asked to place it in one of the trays (e.g., "This is a rabbit, where does that card go?" or just "Rabbit" if the child becomes frustrated with the whole question being asked repeatedly). If the child sorts five cards (or more) correctly, then the child will proceed to the border stage. The border stage implies adding one more if-then condition, where six cards (three red rabbits and three blue boats) out of 12 in total have the added feature of a black border drawn along the edges. The children are instructed to sort the cards according to one dimension (e.g., color) if the card lacks a frame and the other dimension (e.g., shape) if the card has a frame. If the child sorted eight cards correctly, they passed the border stage. The cards were pseudo-shuffled in all stages so the child would not sort cards to the same box more than twice in a row. Each completed stage gave an increased score of 1. That means a completed pre-switch, post-switch, or border stage scored 1, 2, or 3, respectively. The average testing time was four minutes if the child did not perform the border game and ten minutes if the child did.

Measurements of DCCS at two, three, and four years of age are included in the thesis.

The Theory-of-Mind Scale: Wellman & Liu

The Theory-of-Mind Scale (Wellman & Liu, 2004) was used to measure ToM ability. The scale consists of tasks where the child was told different stories with varying degrees of mentalizing demand (Wellman & Liu, 2004). Wellman and Liu (2004) originally used six tasks. Of these six tasks, two measured FB, and one measured understanding of HE; the last was expected to be the most difficult. Only one of the FB tasks, the unexpected content one, was included in the current project. This is in accordance with the final scale in Wellman and Liu (2004), which included five items. Also, the fifth task, tapping understanding of emotional states, was excluded since the children were three years of age at the first testing condition, and their performance was expected to be low. It was also prioritized to keep the number of tasks limited since the total amount of tasks at each testing occasion was large. A cross-sectional study of a sample of typical Swedish children using the Wellman and Liu scale has been published (Sundqvist et al., 2018; the full description of the scale is available in Grape & Sandstig, 2012; Karlsson & Östling, 2012). One result from Sundqvist et al. (2018) showed that the task pertaining to emotion understanding did not follow

the expected developmental trajectory, whereas the other four ones did (i.e., understanding of DD, DB, KA, and CFB). The present thesis included the following four steps:

1. DD (Diverse Desires) – The participants are supposed to understand that others may not have the same preferences as themselves when it comes to food.
2. DB (Diverse Beliefs) – The participants are supposed to realize that others may not have the same beliefs as themselves regarding where a cat can be hiding.
3. KA (Knowledge Access) – The participant is shown something odd about the contents of a box and should recognize that others not shown the contents could not know the contents of that box.
4. CFB (Contents False Belief) – The participant should understand that two things are not always as they seem and that even if the participant knows the fact, others might not.

All tasks have a control (or preliminary) and test questions. A participant passed a task successfully only if the control (and preliminary) question(s) and the test question were answered correctly. Each successfully completed task scored 1, with a maximum score of 4 for the whole scale. The average testing time was ten minutes.

Measurements of ToM at three, four, and five years of age are included in the thesis.

Mental State Talk Test: Picture Book Task

The picture book task was used to measure MST used by the parent. The parent was presented with a plastic binder encompassing ten pictures with emotionally and mentalistic charged situations, such as a child making an angry face towards a peer or two children smiling at a cameraman (at the first testing time, pictures were from Ruffman et al., 2002). The parent was asked to talk about what was happening in the pictures and to switch to the next picture as soon as the child showed that it wanted to turn the page. The experimenter then left the room. New and age-appropriate pictures were used each year. The session was both video-recorded and audiotaped. The dialogue was later transcribed and coded by the authors of this thesis, two more experienced researchers, and seven trained students using a detailed transcription manual. The transcriptions were verbatim,

adding minor details to ease the computerized MST extraction. MST (Ensor & Hughes, 2008) in mothers' language was analyzed for mental state categories, including all references to cognitive terms (e.g., "think" or "know"), emotions (e.g., "happy," "pleased," or "sad") and desires (e.g., "want," "like," or "hope"). The task was scored on the number and proportion of mentalizing words used by the parent, as well as the size of the vocabulary for each mental state category. The experimenter came back into the room after approximately ten minutes. The average interaction time was eight minutes.

Measurements of MST at two and three years of age are included in the thesis.

Off-Site Measurements

The collection of off-site measurements often required reminders. All parents who had not answered their questionnaires got email reminders that they had unanswered questionnaires in their possession almost every week for up to two months. Text message reminders were used occasionally during the same period.

Swedish Early Communicative Development Inventories (SECDI)

A Swedish version of the MacArthur Communicative Development Inventory (MCDI) was used to assess the children's communicative skills. This Swedish Early Communicative Development Inventories, SECDI (SECDI; Berglund & Eriksson, 2000; Eriksson & Berglund, 1999) is based on parental reports and the second version of the SECDI (appropriate for children between 16 and 28 months) was used, which included a measure of productive vocabulary. A short version (431 words in total) was constructed encompassing 13 categories of the complete questionnaire, namely, sound effects and animal sounds, toys, playtime and routines, places to go to, food and beverages, pronouns, words about time, numbers and objects, humans, prepositions and places, verbs, conjunctions and questions, and actions. The Swedish word "tror" was added for the measurement at two years of age but was accidentally removed for the measurement at three years of age. The form was scored on the total number of words the child produced (i.e., the vocabulary as rated by the parent). The questionnaire took approximately 35 minutes to complete.

Measurements of SECDI at two, and three years of age are included in the thesis.

Emotionality, Activity, Sociability, Shyness and Impulsivity (EASI) Temperament Survey

A Swedish version (Hagekull & Bohlin, 1990) of The Emotionality, Activity, Sociability, Shyness, and Impulsivity (EASI) Temperament Survey (Buss & Plomin, 1975, 1984) was used. The questionnaire included 25 statements, with five statements measuring each dimension. All statements were rated on a five-point scale from not at all true (Stämmer inte alls) to very true (Stämmer mycket bra). The questionnaire was answered by parents. The questionnaire took approximately eight minutes to complete.

Measurements of EASI at two, and four years of age are included in the thesis.

Descriptive results

The parent was asked to fill in a short form with questions regarding the participant's name, age, number of siblings, age order amongst siblings, first and second languages in the household, hearing issues and vision issues, and the parents' educational attainment. Education was then ranked on a 7-point scale utilizing the Hollingshead index (Hollingshead, 1975) to get an index of SES (see Table 1). A subsample of the total data collected in the project will be presented in the thesis. A complete listing of the data points collected for each year and test being focused on in the current thesis is found in Table 2.

Table 1 - *Project Sample Demographics*

Measure	2 y.	3 y.	4 y.	5 y.
N	180	149	136	54
Retention rate	100%	83%	91%	40%
% of the baseline sample	100%	83%	76%	30%
Mean age in years (SD)	2.33 (0.07)	3.37 (0.09)	4.36 (0.07)	5.34 (0.05)
% girls	56.1%	56.4%	58.8%	63.0%
% with older siblings	64.4%	61.7%	61.8%	63.0%
% parental dyads avg. BD+	48.9%	48.9%	53.7%	55.6%
% multilingual homes	30.0%	28.9%	28.7%	20.4%
% Swedish as first lang.	87.7%	88.6%	89.0%	92.6%

Note. y. = years old; BD+ = Bachelor's degree or higher; avg. = average; lang = language.

Table 2 - *Test Specification and Sample Size for all Tests and Years Measured.*

Instrument/ Method	Ability/Factor Measured	2 y. 2016	3 y. 2017	4 y. 2018	5 y. 2019	Test time (minutes)*
Hollingshead	SES – Education	180	conf			3
Picture Book	Mental State Talk	180	149			8
DCCS	EF – Shifting	138	149	134		7
The ToM scale	ToM – Verbal		142	134	53	10
SECDI	Lang – Productive	164	130			35
EASI	Temperament	175		126		8

Note. The tests are ordered as they were presented to the participants. Each cell number shows the count of all data collected for that measurement and year.; y. = years old; DCCS = measure of Executive function; ToM = Theory of Mind; SECDI = Swedish Early Communicative Development Inventories; EASI = Emotionality, Activity, Sociability, Shyness and Impulsivity Temperament Survey; conf = participants SES was confirmed using follow-up questions; Blank = not tested; Gray = tested, but not included in the current thesis; Green = tested; Yellow = tested but testing stopped due to Covid-19 pandemic.; * = the number is approximate as testing time varied between participants and the year of measurement.

Summary of Studies

Study I

In *Study I*, the ToM scale was investigated longitudinally in a Swedish context. Previous studies investigating the ToM scale by Wellman and Liu (2004) have mostly been cross-sectional (for a narrative review, see Pava, 2019). Only two previous studies (Peterson & Wellman, 2019; Wellman et al., 2011a) have been longitudinal, and only one has measured the ToM scale at three different time points (Wellman et al., 2011a). It has not been common practice to evaluate the ToM scale properly, and the results have been mixed when the scale has been evaluated using all measures included in the original study by Wellman and Liu. The field´s current state suggests that the ToM scale has more reliability in a longitudinal design.

When evaluating the ToM scale using a longitudinal sample, the ambition was to find an explanation for the fluctuating reliability of the scale while also validating the scale in a Swedish context.

Sample

Study I was based on data from the last three measurement points of the Brain, Mind, and Culture: Pathways to Mentalizing, Language, and Reading project. Two exclusion criteria were implemented after disregarding data points recorded from the 30 participants who did not return for testing at three years of age. The exclusion criteria were children not having Swedish as their first language (n = 17), children with hearing or vision impairments (n = 2), and one participant persistently doing the opposite of what was instructed (n = 1). The attrition and exclusion criteria decreased the group by 50 participants compared to the initial sample tested at two years of age. The total number of participants included in the longitudinal study was 130, 118, and 49 for the measurements at 3, 4, and 5 years of age.

Measures

Only measures of performance on the ToM scale were included in this study.

Statistical Analyses

Guttman Scalogram analyses were performed to assess the ToM scales longitudinal stability, cross-sectional scalability (or reproducibility), and reliability (or consistency). Repeated measures ANOVAs were computed to analyze gender differences.

Results

The ToM scale captured development reliably in a Swedish context (see Figure 2[3]). The difference in performance between the different steps of the scale was similar to other studies in similar contexts (Figure 3). Additionally, the participants very rarely diverted from the general developmental order (i.e., DD > DB > KA > CFB). More specifically, longitudinal analysis of the Guttman Scalogram analyses confirmed that the 4-step scale was stable across measurement years from 3–5 years of age (i.e., 86% followed the Guttman scale at all three years of measurement) and similarly stable between sequential years of measurement (i.e., comparisons between 3–4, and 4–5 years of age revealed consistent Guttman pattern at 81% and 92%, respectively). This suggests that performance on the ToM scale is predictable and systematic between years of measurement and across the entire sample. However, the cross-sectional analyses (i.e., when analyzing each year of measurement separately) revealed that the 4-step scale was not always reliable. This means that even though most participants follow the Guttman scale longitudinally, 3- and 4-year-olds do not strictly keep to the 4-step scale at their respective ages.

In light of this, a 3-step scale (i.e., DD < DB < KA) was analyzed. This scale was found to be longitudinally stable across all years of measurement (92%), between 3–4 (86%) and between 4–5 (92%) years of age. When analyzing the ToM scale cross-sectionally within each age group, the 3-step scale was found reliable at four and five years of age but not at three years. This suggests that the 4- and 3-step scales show similar longitudinal stability, with the 3-step scale being cross-sectionally reliable at an earlier age than the 4-step scale[4].

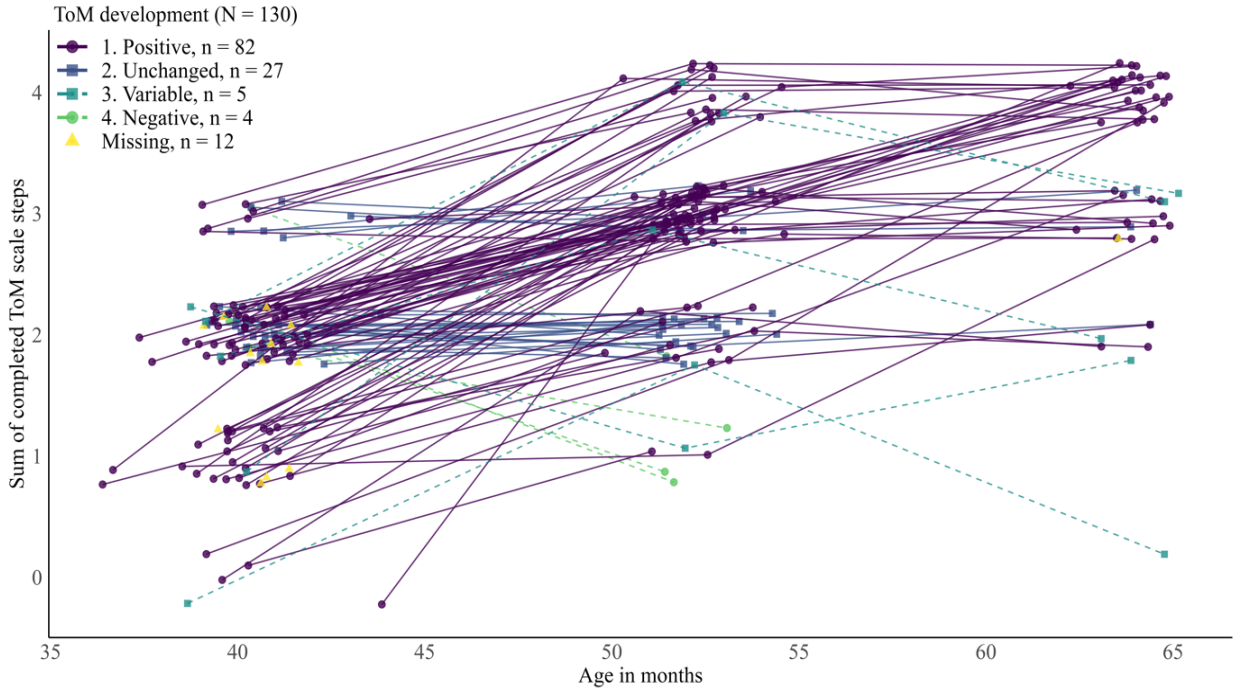
[3] Only 49 participants were measured at five years of age. Twelve participants only measured once (11 at three years of age and one only measured at five years of age).

[4] When preparing for *Study I*, information was gathered about previous results pertaining to the ToM scale. It then became clear that many studies only report Rep and not *I* values. *I* values can be computed given that the researchers reported fail or success rates for each step in the task and number of participants, for each year and samples included. This has been done in many previous studies. However, the *I* value is dependent on Rep and you need to

Furthermore, gender analyses for a smaller sample size analysis over all three years of measurement revealed a general effect of gender $F(2,92) = 5.757$, $p = .026$, partial $\eta^2 = .104$, 95% CI [0.05, 0.78] (but not an interaction with age $F(2,92) = 1.050$, $p = .354$, partial $\eta^2 = .022$; Figure 4A). Nonetheless, no average gender differences $F(1,116) = 1.457$, $p = .230$, partial $\eta^2 = .012$, 95% CI [-0.09, 0.36], or interactions with age $F(1,116) = 0.251$, $p = .618$, partial $\eta^2 = .002$, were found for summation of the ToM scale scores over the first two years of measurement (see Figure 4B).

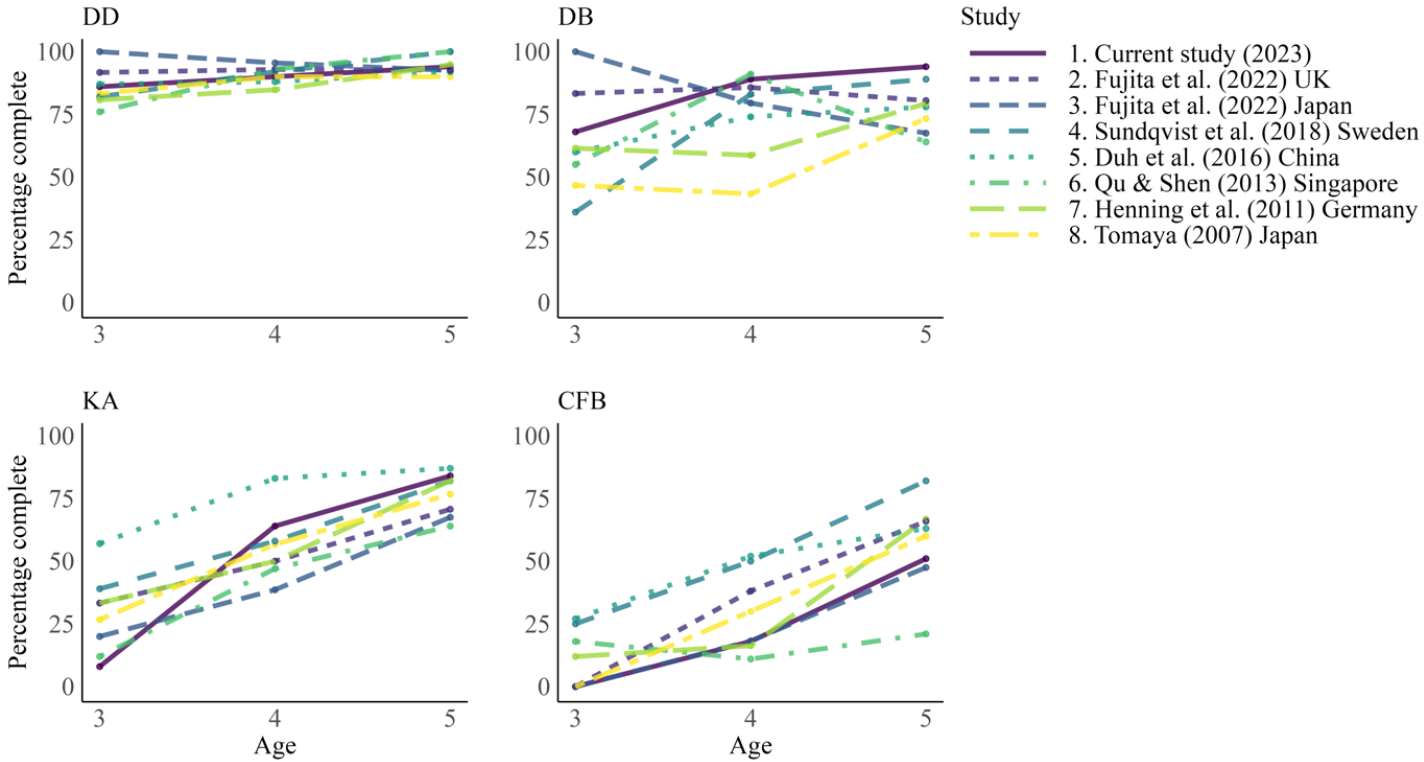
report Rep with three decimal points for the *I* to be computed reliably. It seems common practice to only report *I* only using two decimal points, probably because of APA standards, making the computation very unreliable. It would be beneficial for future studies to report Rep (and *I*) using three decimal points.

Figure 2 - *Longitudinal Trajectories of ToM, Separated by Their Developmental Patterns.*



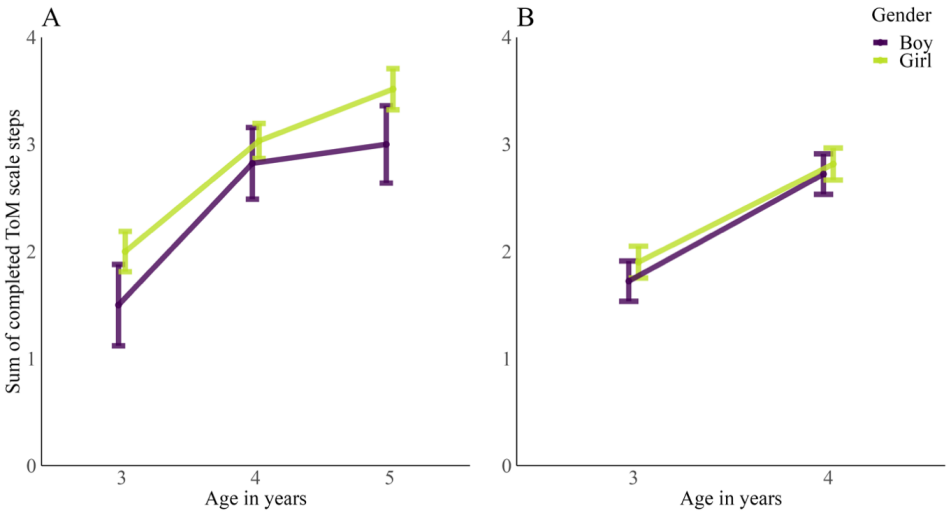
Note. Participants are grouped (and the lines are colored) based on change scores between the measurement at three and four years of age and four and five years of age, respectively. Positive development refers to participants with a positive change score between at least two measurements and no negative change scores. Negative development was represented by participants who had at least one negative change score and no positive change scores. Unchanged development had no positive and no negative change scores. Variable development had one positive and one negative change score. ToM scores are jittered with a random value between -.2 and +.2 to make data points better visible (i.e., the only possible scores on the scale are 0, 1, 2, 3, and 4).

Figure 3 - *Average Completion of ToM Scale Steps in 3–5-Year-Olds.*



Note. The Figure depicts results from seven previous studies reporting performance on the ToM scale in children aged 3–5 for each age and step. When comparing the current study to the other results presented, the sample had comparable performance on Diverse Desires (DD) at all years. Performance on Diverse Beliefs (DB) and Knowledge Access (KA) started low at three years of age but was high in later years. Finally, the current sample had relatively low performance on Content False Belief (CFB) at all ages measured.

Figure 4 - *Difference Between the Larger Sample over Three Years and the Smaller Sample over Two Years for Gender.*



Note. ToM = Theory of mind. Figure (A) depicts the difference between the smaller sample (n = 49) that participated in all years measured and (B) the larger sample (n = 118) that only participated in the first two years. The analysis of the smaller sample showed a significant gender difference where girls performed better than boys. That gender difference is not present in the larger sample. Error bars show 95 % confidence intervals.

Study II

Study II aimed to investigate the role of various temperament dimensions in ToM development controlling for EF, language, and SES. These factors are commonly included as covariates when investigating contributions of other influences, as they are generally related to ToM.

The most common finding concerning the relation between ToM and temperament is that being shy (or socially withdrawn) is solely (or in symphony with other types of temperament) related to ToM (e.g., LaBounty et al., 2017; Lane et al., 2013; Mink et al., 2014; Wellman et al., 2011b). It has also been found that increased inhibition ability might be related to ToM development (e.g., Longobardi et al., 2017; Suway et al., 2012). However, most previous findings have been between temperament and FB tasks. Only one previous study has evaluated the ToM scale by Wellman and Liu in relation to many different temperament dimensions (Henning et al., 2011) and not only a single temperament measure (i.e., a measure of aggression by Song et al., 2016). Henning et al. (2011) reported that the Activity dimension of the EASI form was the closest to significantly correlated to the ToM scale across their four groups aged 3–6 (n = 146). None of the correlation- or regression analyses they performed showed a significant relation between temperament and the ToM scale. More studies investigating the relation between earlier temperament and later ToM development might give new insights.

Sample

Study II was based on data from the first three measurements of the Brain, Mind, and Culture: Pathways to Mentalizing, Language, and Reading project. Exclusion criteria were children with hearing or vision impairments (n = 5); those who did not return for testing at three years of age (n = 29), did not have Swedish as their first language at home (n = 15), and participant-related test difficulties (e.g., not filling in temperament forms correctly, or persistently doing the opposite of what was instructed; n = 10). The exclusion criteria decreased the group with 59 participants compared to the initial sample tested at two years of age.

The final sample included 121, 121, and 111 for ages 2, 3, and 4, respectively.

Measures

Analyses were conducted on measures of temperament (EASI) at two and four years of age, ToM at three and four years of age, productive language (SECDI) at two or three years of age, EF (DCCS) at two, three, or four years of age, and SES (as measured by averaged parental education) measured at two years of age.

Statistical Analyses

The development of ToM scores was examined using Guttman scalogram analyses. Spearman correlations were performed to investigate relationships within and between measures. Structural equation modeling applying path analyses was used to investigate predictive and concurrent relations with the ToM scale.

The structural equation model was evaluated using goodness-of-fit (GOF) measures. GOF measures that are generally recommended were applied. Specifically, the GOF measures and cut-offs used were the Comparative fit index (CFI) > .9, Tucker Lewis Index (TLI) > .9, standardized root mean square residual (SRMR) < .09, a Root mean squared error of approximation (RMSEA) < .05, and, given our small sample size, a not significant Chi2. Robust alternatives to the GOF measures that better handle small sample sizes were chosen when available and marked by a raised letter r (i.e., ^{r}).

Results

When considering the zero-order correlations (Table 3), seven were significantly related to ToM at three and four years of age. ToM at three years of age was correlated with ToM at four years of age (r = .30, p = .001) and SES (r = .23, p = .012). ToM at four years of age was correlated with language (r = .24, p = .033) and the temperament dimension Activity (r = -.23, p = .014) at two years of age, EF (r = .24, p = .010) and language (r = .21, p = .011) at three years of age, and EF (r = .28, p = .004) at four years of age. To summarize, all measured concepts except the temperament dimensions of Shyness and Emotionality were associated with ToM at one or two measurements.

The path analysis performed (see Figure 5) was found to have an appropriate fit (Chi2^{r} (30) = 36.929, Chi2 p^{r} = .179, CFI^{r} = .900, TLI^{r} = .817, SRMR = .052, RMSEA^{r} = .042, RMSEA 95% CI^{r} = [0, .08]). The analysis revealed a few significant findings. Regarding ToM and temperament, Shyness at two years of age had a negative relation to ToM at four (Est = -0.348, p = .011, 95% CI = [-

0.62, -0.08]) years of age. This suggests that Shyness is a factor in ToM development even when controlling for ToM the previous year.

With regards to other relations included in the path analyses, two control variables were found to be positively significant, namely SES (Est = 0.134, p = .034, 95% CI = [0.01, 0.26]) and EF (Est = 0.260, p = .047, 95% CI = [0.00, 0.52]) at two years of age were significant in relation to ToM at three years of age. Importantly, neither EF at two years of age (or at any age for that matter) nor SES was associated with ToM at four years of age (i.e., when previous ToM was included in the analysis). Additionally, the child's productive language had no significant association with ToM at any age.

In addition, four within-measure regressions were significant and positive. ToM from three to four years of age (Est = 0.295, p = .007, 95% CI = [0.08, 0.51]), EF from two to three years of age (Est = 0.216, p = .028, 95% CI = [0.02, 0.41]) and from three to four years of age (Est = 0.337, p < .001, 95% CI = [0.15, 0.52]), and language from two to three years of age (Est = 0.220, p < .001, 95% CI = [0.13, 0.31]).

Pseudo-R^2 for the model was 10% for ToM performance at three years of age and 23% for ToM performance at four.

In sum, there were significant relations between ToM and temperament even when controlling for many other variables previously associated with ToM[5]. Interestingly, none of the other control variables were significantly associated with ToM at any measurement year when controlling for previous ToM.

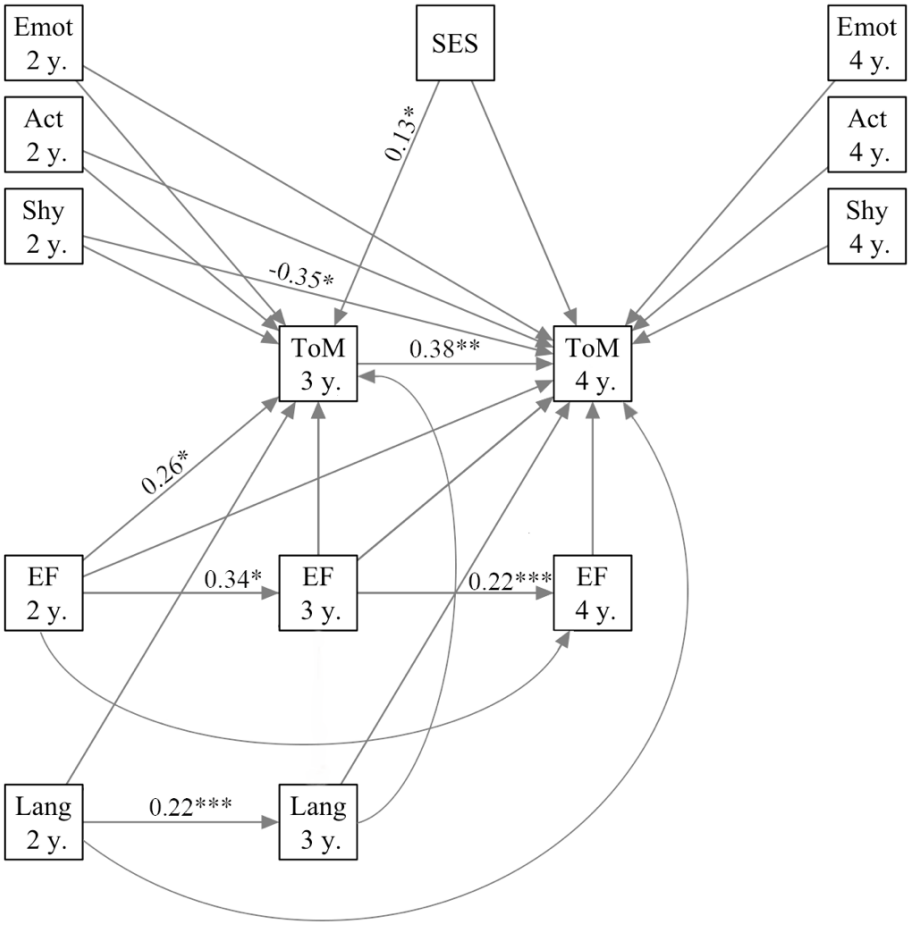
[5] For the sake of completeness, I present an extra path analysis in *Study II*. A reanalysis of the path analysis above was performed with the only difference that the ToM scale only included DD, DB and KA. However, when considering the goodness of fit measures for the 3-step scale, CFI and TLI was found to be low (Chi2^r (30) = 37.798, Chi2 p^r = .155, CFI^r = .881, TLI^r = .782, SRMR = .055, RMSEA^r = .047, RMSEA 95% CI^r = [0, .09]). Therefore, the manuscript concentrated on results from the 4-step scale in *Study II* but still included findings on the 3-step scale. Nonetheless, three regressions were significant in the analysis with the 3-step scale. The first two were in relation to ToM at four years of age, namely ToM at three (Est = 0.218, p = .017) years of age, and Shyness at two years of age (Est = -0.271, p = .009). The last significant regression was between ToM at three years of age and SES (Est = .145, p = .017) In sum, the only other variables included in our analysis associated with ToM, apart from ToM itself, was Shyness and SES.

Table 3 - *Mean, Standard deviation, the Range for All Included Variables, and Correlation Between All Variables In Study II.*

Variable	M (SD)	Range	1[a]	2[a]	3[a]	4[a]	5[a]	6	7	8	9	10	11	12	13	14
1. ToM 3y.[a]	1.84 (0.66)	0.00–3.00		110	100	121	108	114	98	120	118	121	116	104	103	104
2. ToM 4 y.[a]	2.77 (0.80)	1.00–4.00	.30**		92	110	108	103	88	109	107	110	105	102	101	102
3. EF 2 y.[a]	0.66 (0.48)	0.00–1.00	.17	.12		100	90	95	80	100	98	100	95	85	84	85
4. EF 3 y.[a]	1.35 (0.51)	0.00–2.00	.11	.24*	.20*		108	114	98	120	118	121	116	104	103	104
5. EF 4 y.[a]	1.89 (0.54)	1.00–3.00	.17	.28**	.15	.35***		101	86	107	105	108	103	100	99	100
6. Lang 2 y.[b]	2.13 (0.97)	0.12–4.18	.07	.21*	.19	.18	.21*		92	113	111	114	110	98	97	98
7. Lang 3 y.[b]	3.64 (0.43)	2.35–4.22	.02	.20	.18	.05	.14	.45***		97	96	98	93	88	87	88
8. Avg. Ed.	5.58 (0.97)	3.00–7.00	.23*	.12	-.07	.07	.15	-.10	-.10		118	120	115	103	102	103
9. Emot 2 y.	3.42 (0.62)	2.20–5.00	.01	-.04	.05	.05	-.10	.08	.17	-.17		118	113	102	101	102
10. Act 2 y.	3.90 (0.66)	2.60–5.00	-.17	-.23*	-.01	-.15	-.07	-.07	.05	-.16	.21		116	104	103	104
11. Shy 2 y.	2.10 (0.65)	1.00–4.00	.02	-.09	-.11	-.01	-.20*	-.11	.11	.11	.04	-.44***		99	98	99
12. Emot 4 y.	3.45 (0.75)	1.80–5.00	.08	.05	.22*	.13	-.05	.10	.18	-.03	.38***	.06	.09		103	104
13. Act 4 y.	3.68 (0.72)	2.00–5.00	.04	-.14	.00	-.05	-.04	-.04	.04	-.25*	.26**	.65***	-.37***	.25*		103
14. Shy 4 y.[c]	2.19 (0.91)	1.00–4.50	-.07	.04	-.15	.04	-.11	-.15	-.09	.26**	-.05	-.17	.61***	.08	-.33***	

Note. The upper half presents n for each pairwise correlation. The lower half presents correlation coefficients. Significant values are in bold.; y. = years old. ToM = Theory of mind; EF = Executive function; Lang = Language; Avg. Ed. = Averaged parental education; Emot = Emotionality; Act = Activity; Shy = Shyness; [a] = Spearman correlation; [b] = score divided by 100; [c] = score calculated on four items. * = $p <$ .05; ** = $p <$ .01; *** = $p <$ .001.

Figure 5 - *The Path Model Used in Study II to Analyze the Data with Significant Paths Marked.*



Note. All numerical estimates are placed above the regression lines to which they belong.; y.= years old; ToM = Theory of Mind; EF = executive functions; Lang = productive language; SES = socioeconomic status as measured by average parental education; Emot = Emotionality; Act = Activity; Shy = Shyness.

Study III

Study III focused on the social aspects related to ToM, with a prime interest in the contributions of MST. Both proportional and absolute (or frequency) values of MST have been presented previously. Still, only absolute values have received meta-analytic support, maybe because few previous studies report proportional values (Tompkins et al., 2018). Investigations into the relations between MST and ToM have only been done once with two-year-old children (Ensor et al., 2014). More studies investigating the relation between early MST and later ToM development might give new insights.

Similar to *Study II*, to make the analysis more informative and to map the interplay between other explanatory variables, language, EF, number of siblings, and SES were included as covariates.

Study III aimed to investigate how the parents' MST was related to the child's ToM understanding and to describe the mentalizing vocabulary used by Swedish parents in conversation with their children. Both absolute and proportional values of MST in relation to later ToM abilities were included. Additionally, one previous study has analyzed emotional vocabulary size in relation to emotional understanding (Martin & Green, 2005); however, it seems as if an analysis of parental MST vocabulary size has never been carried out in relation to ToM. As parental MST vocabulary size is a potential factor in developing the child's ToM, this aspect of verbal communication was also included in the analyses.

Sample

Study III was based on data from all four measurement points of the Brain, Mind, and Culture: Pathways to Mentalizing, Language, and Reading project. Exclusion criteria were children not having Swedish as their first language (n = 15), inaudible speech or parents speaking another language than Swedish during MST (n = 8), and children with hearing or vision impairments (n = 2). Finally, families that did not have the same parent present at all measurements were excluded to ease the interpretation of the results (n = 30). One additional child was excluded since it did the opposite of what was instructed when tested at three years of age (n = 1).

After attrition and applying the exclusion criteria mentioned above, 80 participants (52 girls) were included at each measurement year; however, testing at five years of age was halted before completion because of the Covid-19

pandemic in April 2020. Therefore, only 32 participants (20 girls) were tested at five years of age.

Measures

Regression analyses were conducted on measures of MST at two and three years of age and ToM at 3, 4, and 5 years of age. Language (SECDI), EF (DCCS), SES (as measured by averaged parental education), and number of siblings at two years of age were included as control variables.

Statistical Analyses

Spearman correlations were performed to investigate relationships within and between measures (Table 4). Latent growth curve model (LGCM) analysis was used to investigate predictive and concurrent relations with the ToM scale (Figure 6 and Table 5). Absolute frequency, proportions, and vocabulary size estimations were analyzed in separate LGCMs. The same GOF measures and cut-offs as in *Study II* were used to evaluate the models included in *Study III*.

Results

The basic LGCM analyses showed that the performance on the scale progressed as expected, with a stable increase at successive years of measurement.

Six significant results were found regarding the uncontrolled (i.e., zero-order) correlations between ToM and MST (Table 4). Two significant negative correlations were between ToM at three years of age and absolute frequency ($r = -.235$, $p = .039$) and proportions ($r = -.229$, $p = .043$) of cognition words at two years of age. Two significant positive correlations with ToM were found at four years of age. The first was related to the proportions of cognition words at three years of age ($r = .263$, $p = .018$). The other significant finding associated with ToM at four years of age was measurements of cognitive vocabulary size ($r = .314$, $p = .005$). Lastly, one significant positive correlation was found between ToM at five years of age and emotional vocabulary size at three years of age ($r = .557$, $p < .001$). In sum (when considering the correlation results), parents' propensity to use cognitive words analyzed as absolute frequency and proportions was associated with the child's later ToM ability. Also, parent's earlier ability to vary their cognitive and emotional vocabulary was related to the child's later ToM.

Three parallel LGCMs revealed, in total, five significant results between children's ToM and parental MST (summarized in Figure 6 and Table 5). The first three were related to children's ToM ability at four years of age. There were negative associations with proportions of parental cognition words at two years of age (Est = -0.10, p = .033, 95% CI = [-0.19, -0.01]) and proportional desire at three years of age (Est = -0.44, p = .001, 95% CI = [-0.75, -0.13]). Furthermore, there were positive relations with proportions of parental cognition words at three years of age (Est = 0.11, p = .029, 95% CI = [0.01, 0.21]). The last two findings were related to the developmental ToM trajectory (or change), namely to the absolute frequency of parental cognition words at two years of age (Est = 0.01, p = .035, 95% CI = [0.00, 0.03]) and parental emotion vocabulary at three years of age (Est = 0.08, p = .027, 95% CI = [0.01, 0.16]). In sum, parents' propensity to use proportionally more cognitive words at two years of age and proportionally more desire words at three years of age were associated with children having lower ToM. However, parents' propensity to use proportionally more cognitive words at three years of age was associated with their children having better ToM. Furthermore, parents' propensity to use more cognitive words and a more varied cognitive vocabulary also had a positive association with the children's rate of ToM development. At the same time, parents' propensity to use proportionally fewer cognition words at two years of age and desire words at three years of age was associated with the child having a better ToM ability.

The MST and ToM associations aside, the child's productive language and SES were significantly related to ToM. Specifically, the child's productive vocabulary was associated with its ToM ability regardless of the type of MST included in the analysis. Additionally, SES was related to ToM ability for analyses with analyses that included absolute and proportional parental MST but not the analysis that included parental MST vocabulary size. Importantly, neither productive language nor SES was significantly related to individual rate of change of ToM development.

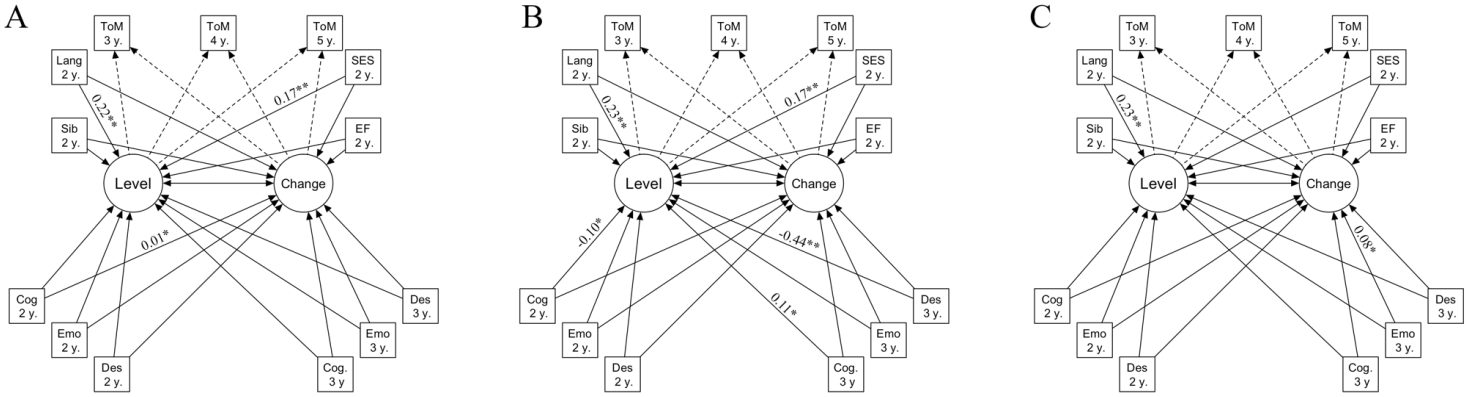
Pseudo R^2 for ToM level and rate of change of ToM was for the absolute frequency analysis 39% and 44%, for the proportional analysis 56% and 74%, and for the vocabulary size analysis 53% and 63%, respectively.

Table 4 - *All Spearman Correlations in Study III Excluding Correlations Between MST Variables.*

Measure	ToM 3y.	ToM 4y.	ToM 5y.	EF	Lang	SES	Sib
ToM 3y.							
ToM 4y.	.308*						
ToM 5y.	.332	.402*					
EF	.184	.156	.122				
Lang	.124	.285*	.288	.184			
SES	.301*	.121	.336	-.083	.040		
Sib	.022	-.102	-.032	-.037	-.193	.076	
			Absolute frequency				
Cog 2y.	-.235*	-.067	.206	-.177	.103	-.088	-.122
Emo 2y.	.019	-.017	.234	.109	-.003	-.029	-.020
Des 2y.	-.023	-.036	.235	-.094	.046	-.140	-.059
Cog 3y.	-.094	.213	.145	-.052	.105	.039	-.096
Emo 3y.	-.021	.054	.271	-.028	.099	.192	-.018
Des 3y.	.026	.089	.095	-.086	.227	.082	.104
			Proportion				
Cog 2y.	-.229*	-.038	.087	-.151	.027	-.007	-.191
Emo 2y.	.037	-.049	.306	.143	.062	.068	.095
Des 2y.	.034	.004	.150	-.001	.086	-.054	-.103
Cog 3y.	-.018	.263*	.064	-.013	-.010	-.016	-.013
Emo 3y.	.014	.124	.360	.106	.015	.154	.028
Des 3y.	-.015	.019	-.130	-.055	.179	.069	.144
			Vocabulary size				
Cog 2y.	-.078	-.171	-.062	-.178	-.121	-.092	-.143
Emo 2y.	-.094	.026	.342	.073	.091	.015	.057
Des 2y.	.000	-.162	.155	.021	.039	.015	.076
Cog 3y.	.043	.314**	.084	-.113	.089	.068	-.172
Emo 3y.	.089	.175	.557***	.046	.201	.191	-.185
Des 3y.	-.009	.209	.149	-.071	.063	-.017	-.036

Note. ToM = theory of mind; y. = years of age; Lang = productive language; SES = socioeconomic status; Sib = number of older siblings; EF = executive function; Des = desire; Emo = emotion; Cog = cognition; * p <.05; ** = p < .01; p < .001.

Figure 6 - *Latent Growth Curve Models Used in Study III to Analyze Theory of Mind Development.*



Note. Subfigure A = Absolute frequency of MST; Subfigure B = Proportion of MST; Subfigure C = Vocabulary size of MST; ToM = theory of mind; y. = years of age; Lang = productive language; SES = socioeconomic status; Sib = the number of older siblings; EF = executive function; Des = desire; Emo = emotion; Cog = cognition; Level = variable latent intercept (or individual ToM ability at four years of age); Change = variable latent slope (or change in ToM from 3–5 years of age). Solid lines indicate regressions and dashed lines indicate fixed parameters relating to estimating the level and change.

Table 5 - *Summary of Significant Associations Between ToM and MST in Study III.*

		ToM	
		Level (at 4 years of age)	Change (from 3–5 years of age)
MST type	Absolute frequency		Cognition at 2 y. (+)
	Proportion	Cognition words at 2 y. (-) & Cognition words at 3 y. (+) & Desire words at age 3 y. (-)	
	Vocabulary size		Emotion words at 3 y. (+)

Note. ToM = Theory of Mind; MST = Mental state talk; y. = years old; & = separating significant findings at the same measurement year.; (+) = positive relation; (–) = negative relation; * = $p < .05$; ** = $p < .01$.

General Discussion

This thesis addressed three general research questions: (1) How can ToM development be reliably measured longitudinally in a Swedish context with a specific ToM scale? (2) How do individual factors relate to ToM ability in preschoolers? (3) What social factors are related to ToM development? In the following text, the results of the three studies on which this thesis is based will be discussed. First, the text below will start with a brief summary of the results in relation to the aims. Second, the details of the results will be discussed in relation to previous research. Third, the strengths and limitations of the included studies will be discussed. Lastly, the following sections will discuss the ethical considerations, theoretical implications, and research gaps.

Summary of Results in Relation to Aims

How can ToM development be reliably measured longitudinally in a Swedish context with a specific ToM scale? The current data suggest that some versions of the ToM scale are slightly more reliable than others, depending on the age group measured. Findings from *Study I* suggest longitudinal stability of a 4-step ToM scale in preschool ages. Cross-sectional analyses within individual ages revealed that only 5-year-old children reliably perform in a sequential Guttman pattern of a 4-step ToM scale. Analyses of the 3-step scale showed that it was longitudinally stable. Additionally, the 3-step scale was reliable when investigating ToM development cross-sectionally in individual age groups as young as four years of age.

Consequently, the 3-step and 4-step ToM scales seem to be stable longitudinal measures of development in the preschool ages, with the 3-step scale being cross-sectionally appropriate at a younger age than the 4-step scale.

How do individual factors relate to ToM ability in preschoolers? Study II shows few significant relations between the individual factors (EF, productive language) and ToM. EF was positively associated with ToM at three years of age. Additionally, Shyness at two years of age was negatively associated with ToM at four years of age. However, EF had an unstable association with ToM, with only one significant association found despite three measurements included in the analysis. Additionally, all individual factors (except the temperament dimension Shyness) were absent when the previous ToM was controlled for.

Consequently, the current thesis gives limited support for the association between individual factors measured and ToM development. Importantly, Shyness showed the strongest association with ToM development.

What social factors are related to ToM development? In *Study III*, ToM ability at four years of age was negatively associated with proportional parental use of cognition words at two years of age and positively related to proportional parental use of cognition words at three years of age. These findings present a complicated connection between MST before the age of four and ToM ability at four years of age, and these findings will be reviewed later in the discussion section. Also, proportional parental use of desire words was negatively associated with ToM ability at four years of age. Crucially, absolute parental use of cognition words at two years of age and the size of the parent's emotional vocabulary when talking with their child at three years of age had a positive association with the rate of change of ToM from 3–5 years of age.

Concerning control variables, SES was most often positively related to ToM at four years of age but never to the rate of change in ToM from 3–5 years of age. Additionally, the number of older siblings was not found to be associated with ToM development.

Consequently, the current thesis gives some support for the association between social factors that were measured and ToM development. Specifically, the strongest support was found for parental use of cognitive, emotional, and desire words in communication with their children.

Discussion

The text below will discuss the details of the current thesis. Each headline focuses on a subpart of the three aims guiding the current thesis. The following text intends to situate findings in the present thesis in the relevant research fields. However, where it was found appropriate, discussions of the current findings occasionally include relevant findings in less related research fields.

ToM Scale Reliability

Study I investigated the reliability and scalability of a ToM scale (Wellman & Liu, 2004) in a longitudinal sample of preschoolers aged 3–5 years. The psychometric evaluation was performed using Guttman scalogram analyses (Guttman, 1944). Aggregated analyses were reported across age groups (as performed in Wellman et al., 2011a) in the supplementary information, and cross-sectional analyses within the individual ages (as performed in Peterson & Wellman, 2019) in the main text. In addition, the ToM scale's longitudinal stability was evaluated by investigating if the participants kept to the expected development order at each measurement year.

The expectation that the 4-step scale would be a stable measure of ToM development across the preschool ages was confirmed in the longitudinal stability analyses. Knowing this, the scale was evaluated on subgroups in the dataset. The scale was assessed at each age, and the 4-step scale was cross-sectionally inappropriate at three and four years of age. Therefore, further analysis of the 3-step scale was performed. As with the longitudinal analyses of the 4-step analysis over the total sample, the 3-step scale was stable. The cross-sectional analyses of the 3-step scale were primarily suitable for each age, except at three years of age. This finding can be important when comparing groups, as it is optimal if the ToM scale is appropriate for each group being compared. Analyzing with this type of reasoning can quickly become an issue, for instance, when sample sizes within groups become too small. When considering the previous findings by Wellman et al. (2011a) and Peterson and Wellman (2019) in relation to the findings in *Study I*, the take-home message is best captured in the following statement: the ToM scale might not be consistently appropriate for participants younger than five years of age.

Study I supports and extends the longitudinal findings, including samples from other countries and cultures (Wellman et al., 2011a) and older age groups (Osterhaus et al., 2022; Peterson & Wellman, 2019). For instance, Wellman,

Fang et al. (2011) never reported a longitudinal stability analysis. Instead, they report the scalability and reliability of a 4-step scale (i.e., DD>KA>FB>HE) across two time points for all included groups. Still, their analyses show that their ToM scale had acceptable scalability and reliability for their included groups. With regards to Peterson and Wellman's (2019) results, they only report cross-sectional analyses of scalability (but not reliability) of the ToM scale for the individual groups they included (i.e., typically developing children, children with autism spectrum disorder, and deaf children with hearing parents). Almost all groups were found to have high scalability, except the group with deaf children. Therefore, it is likely that most groups in their study also had acceptable reliability.

Moreover, the longitudinal sample was used to extend the previous findings (i.e., Sundqvist et al., 2018) regarding ToM development in Swedish preschoolers. Both girls and boys were found to conform similarly to the expected Guttman pattern, with girls having a slight numerical advantage. Additionally, boys and girls showed a comparable and stable advancement in ToM understanding during the early preschool years. Additionally, the ToM development observed was close to studies conducted in similar countries (for a narrative review, see Pava, 2019). Therefore, the findings align with the anticipated result that the ToM scale is sensitive to age-related differences in ToM ability during preschool (Wellman et al., 2006; Wellman & Liu, 2004). Also, the 4-step ToM scale shows sensitivity to individual differences in ToM ability for children aged 3–5, with ceiling effects becoming very likely in older ages (Devine, 2021; Wellman & Liu, 2004).

ToM Development

In *Study II* and *Study III,* the results reveal that when previous ToM scale scores were included in an analysis, few other factors were associated with ToM (see the *Temperament in Relation to ToM* and *Parental MST in Relation to Children's ToM* sections for the exceptions). This suggests that EF, language, family composition, and the SES measures included are less related to ToM than parental MST and child temperament in the present sample. Importantly, the findings do not suggest that the classically relevant measures are irrelevant, merely less relevant relative to some measures of MST and temperament.

In line with the findings, a recent book chapter by Devine (2021) presents a meta-analysis of ToM development from 24–176.5 months (i.e., 2–14.7 years) of age with a focus on rank-order development. The results support the

suggestion that ToM development, much like personality and intelligence, is non-trivial and that early ToM is indicative of a long-term outlook for development. Devine (2021) further extends these findings by reviewing research that found ToM in early childhood (i.e., age 6) to have a unique association with social competence in middle childhood (i.e., age 10), above and beyond that of EF and language comprehension (Devine et al., 2016). Therefore, Devine (2021) suggests that a substantial portion of individual differences in ToM may be related to differences in propensity (e.g., willingness to contemplate about, and sensitivity towards, other´s perspectives) and fluency (e.g., insight into the associations between mental states and various contexts).

The current findings, especially those reported in *Study I*, support the findings showing that ToM development is stable and gradual when measured yearly. However, how ToM measurements were collected naturally limits our conclusions to that timescale, making it impossible to investigate the stability of ToM development day-to-day or week-by-week. Fortunately, there are some findings from investigations of ToM ability many times during one year. For instance, Baker et al. (2016) measured EFB and CFB monthly in a small sample of 34–64-month-old children. Using Bayesian change point modeling, they found weak support for stable and gradual, or "sudden insights" in ToM development when measured monthly. Instead, they suggest a relatively stable increase in ToM ability over long periods of time, but the ability in the short term could be very variable. Baker et al. (2016) suggest that there might be unidentified factors related to ToM development in shorter intervals than one year, with some instability still present after one year. The findings in *Study I* support their conclusion, as even the current thesis includes subjects with a variable ToM development on the year-to-year timescale (i.e., 5 out of 49 participants measured at all three occasions). Therefore, more research is needed to bring forth these currently occluded factors.

Language in Relation to ToM

Relations between productive language and ToM development were (mainly) investigated in *Study II*. However, even if *Study II* included repeated measurements of productive language, significant associations were only found with ToM in *Study III* and consistently with the level of ToM ability at four years of age, regardless of whether the analysis was performed with absolute frequency, proportions, or vocabulary size MST values. The findings suggest that productive language is only inconsistently associated with ToM and never

when previous ToM is controlled for. Nevertheless, results align with previous studies involving similar measurements when not controlling for earlier ToM (e.g., Brooks & Meltzoff, 2015; Farrar & Maag, 2002; Watson et al., 2001).

Production and Comprehension

The included language measure (i.e., SECDI, the Swedish version of the MCDI) excluded grammar items; consequently, a productive vocabulary measure remained. The findings by de Mulder et al. (2019) showed that sentence comprehension around four years of age was related to measures of ToM eight months later, even when controlling for earlier ToM, age, syntax ability, and mental vocabulary. They also found that comprehension vocabulary was associated with later ToM when only controlling for earlier ToM and syntax ability. Similarly, Devine et al. (2016) found that comprehension at six years of age (when controlling for age, SES, EF, and teacher-rated social competence) was associated with ToM at ten years of age. The current thesis does not align with the findings of de Mulder et al. or Devine et al. For instance, a crucial difference between *Study II* and *Study III* and de Mulders et al.'s and Devine et al.'s result is that the current thesis never found that productive language was associated with ToM when previous ToM was controlled for. This discrepancy might be because productive language is less related to ToM than comprehension. However, Milligan et al. (2007) never made this comparison, as productive vocabulary was not a part of the meta-analysis. Nonetheless, the information gathered in the current thesis, compared with previous research into comprehension, suggests a slight advantage for comprehension over production.

A related and important question is how productive language relates to comprehension since both characterize communicating and interacting with others. Bottema-Beutel et al. (2019) found longitudinal correlations for productive and comprehension measures when investigating 8–32-month-old children, with early comprehension vocabulary predicting later productive vocabulary and vice versa. To extend these findings, the relationship between language comprehension and production has been investigated internationally in a large cross-sectional survey study by Bornstein and Hendricks (2012). They reported, using a subsample (n = 38845) of children older than one but younger than five years of age, positive correlations between production and comprehension for most included countries and ages but found weak correlations within most countries (weighted mean r = .22). This suggests that production and comprehension are abilities that are only vaguely indicative of each other.

Considering these previous results and the language measure used in the current thesis, the association between productive language ability and ToM might bear relations to ToM outside of comprehension. The reasoning is that comprehension is weakly related to production, and production is associated with ToM. Therefore, a child with a vast productive vocabulary might better aggregate experiences by using a vocabulary better suited for the social situations the child experiences at home and preschool. Even though the results suggest that productive language seems less relevant when including previous ToM, future investigations including productive vocabulary combined with measures of comprehension and grammar (and a general measure of language) could give new insights (echoing Farrar et al., 2017; Milligan et al., 2007).

Insights from Neuroscience: Studies on Adults

The relationship between language and ToM might be illuminated by including neuroscientific findings. One meta-analysis by Schurz et al. (2021) investigated the overlap between neuronal architecture related to ToM, empathy, and many other cognitive functions among adults. As a reference, FB tasks were found to map to more cognitive brain clusters (e.g., cortical midline, temporoparietal areas, and medial prefrontal cortex). Observing emotions (i.e., looking at emotional faces in the Reading Mind in the Eyes task; Baron-Cohen et al., 2001; Baron-Cohen et al., 1997) were found to map to more emotional brain clusters (e.g., insulae, supramarginal gyri, and right temporal pole). The most crucial insight is that the FB and Reading Mind in the Eyes tasks can be found on different extremes of the cognitive-affective spectrum, echoing Wimmer and Perner (1983) and Baron-Cohen et al. (1997). With that finding in mind, they also report that language areas were more related to ToM connected to affect (e.g., Reading the mind in the eyes task) compared to ToM focusing on cognition (e.g., CFB or EFB), suggesting that language is more important to the former, compared to the latter in adults. More research is needed to elaborate on Schurz et al.'s (2021) findings. In particular, the developmental trajectory from childhood to adult ages.

EF in Relation to ToM

Study II was designed to investigate concurrent and predictive relations between EF and ToM development. When relations between EF and ToM were exclusively investigated, some significant findings were found: EF at two years of age is related to ToM development at three years of age (as seen in *Study II*).

Multiple (or latent) measures of EF would have given us a richer approximation of the children's EF abilities and, perhaps, different conclusions. Still, the chosen measure of EF was never significant when controlling for other factors. Therefore, the results do not support EF being amongst the most potent factors for ToM development. Similarly, the lack of significant findings with EF does not support EF and ToM development being linked or developed in parallel through the mediation of social interaction (Moses & Tahiroglu, 2010). Nonetheless, ToM cannot exist without the support of EF, as thinking about other's perspectives in various situations and keeping relational information in mind implies taxing EF.

Zero-Order or Controlled Analyses

The lack of associations between EF and ToM seems to contradict the findings presented in an intervention study on 3–4-year-old children (n=44) by Kloo and Perner (2003), where practice on EF (precisely the DCCS task that was used as an EF measure in the current thesis) improved FB ability, and not vice versa. The results also contradict the meta-analysis by Devine and Hughes (2014), finding associations between EF and FB. However, that meta-analysis was performed on zero-order correlations and, for the most part, did not control for related variables. Focusing on the zero-order correlations in this thesis, the results align perfectly with Devine and Hughes's findings. A subset of their meta-analysis, including 48 studies (n = 3584), allowed for partial correlations, controlling for age and verbal ability. There was still a small to medium relation between EF and FB in this group of studies. Given the intricate relation between ToM and other factors, controlling for at least some variables is appropriate to understand how ToM develops. Verbal ability, maternal education, and previous ToM ability were included in the path analysis in *Study II*. In that analysis, EF was found to be of no crucial importance for ToM in the preschool years when evaluated in symphony with these other factors.

The question is now, might there be an explanation for this lack of associations between ToM and EF? Perner and Lang (1999) suggested that the mental strain of being tested on many tests may dilute the relation between EF and ToM/FB. As the test battery at each measurement year included many different tests, that factor might be relevant. However, Devine and Hughes (2014) did not find support for Perner and Lang's reasoning with their much larger inclusion of studies (i.e., 100 different effect sizes) and more systematic analysis. One aspect that has stayed the same between Perner and Lang's study

and the study by Devine and Hughes is the heterogeneity of results when analyzing relations between EF and ToM. When evaluating the heterogeneity further, Devine and Hughes (2014) found that larger sample sizes may negatively affect the relation between ToM and EF. They explain that more extensive studies often include many different experimenters, thereby introducing more variation to the test procedure than when the same experimenter tests all participants in smaller studies. They also suggest an alternative interpretation by stating that larger samples may be more representative than smaller ones. The median sample size in their meta-analysis was 68, with an interquartile range of 42 (i.e., 50 % of the sample sizes previously reported are roughly between 47 and 89). When considering Devine and Hughes' definition, *Study II* can be described as an average study considering sample size (i.e., that should find a significant relation between EF and ToM). Moreover, all participants were tested by the same experimenter (i.e., no extra random variation was introduced by including many different experimenters). In this light, and with Devine and Hughes's reasoning in mind, *Study II* and *Study III* did present the expected result (i.e., a relation between ToM and EF) when considering study design and zero-order correlations. However, as stated earlier, the role of EF might not be as relevant when controlling for many ToM-related factors as done in other (and often more advanced) statistical analyses.

ToM and EF Load

Discussing basic test design when considering the results is important, as some tests might tax EF more than others. Setoh et al. (2016) presented a study investigating EF and language loads for solving FB tasks. They reported that children are often found unable to pass FB tests before four years of age, but if the EF and language load tests are lowered (by using control questions), then a child as young as two and a half years of age could be successful. However, Setoh et al.'s criticism might not entirely apply to the ToM scale. All steps in the ToM scale have been designed to put as equal demands as possible on EF and language within steps and differences in demands between steps; Wellman, 2014) by using control questions and visual aids, as suggested by Setoh et al. (2016). Additionally, no objects changed containers in the FB task (the fourth step, CFB).

Importantly, there is a difference between the CFB task and the task that Setoh et al. (2016) investigated. Namely, they used a task where objects change containers to create a situation with FB (i.e., an EFB task). An unexpected

contents task was included in the current thesis, where the child realized that a band-aid box contained something unexpected, namely a nail. Setoh et al. (2016) removed the object from the box it had moved to in the low-demand setting, and I did not remove the nail. Therefore, the crux remains in the current project is that the item was showed in to the participants in the band-aid box remained in the box when an agent appeared to guess what was in the box. The ToM scale might tax EF slightly more than Setoh et al.'s (2016) design. This might be because the test procedure still has some EF load. To test this, additional analyses were performed to investigate this possibility by correlating the EF measure and individual ToM scale steps in *Study II,* and no significant correlations with CFB were found. The significant results that were found were predictive and concurrent correlations with KA. Additionally, the concept of low-demand FB tasks being able to capture "early" FB has been criticized and hard to replicate (according to Wang et al., 2019). Still, future research is needed on the topic.

Temperament in Relation to ToM

The analyses in *Study II* only revealed limited support for temperament being a significant predictive factor for ToM development. The only significant result was that Shyness had a negative association with ToM. However, before getting into details of the results, a theory suggesting that children might be influencing their social environment should be briefly outlined.

Social Tendencies

The Social tendencies theory by Lane and Bowman (2021) generally states that children affect the social situations they encounter through their way of approaching them. Lane and Bowman (2021) reviewed 37 previous findings and reported that there seemed to be some systematic patterns in the previous literature. For example, Lane and Bowman (2021) conclude that Activity appears to be negatively related to ToM. Regarding Shyness, their review of previous findings suggests that shy and socially observant behavior is positively associated with ToM. The results align with the general claim in the Social tendencies theory but not with the specifics.

Shyness

The results are in line with the general conclusion that out of all the temperament dimensions, the social-withdrawal (or Shyness) has been the most frequent significant predictor of FB and ToM scale scores (for a review, see Lane &

Bowman, 2021). However, the findings did not fit the specifics of the relation between Shyness and ToM, namely, that more Shyness is related to better ToM, as the results suggest that less Shyness at two years of age was related to better ToM at four years of age.

It is not apparent why this finding was obtained. There is a discussion about positive (e.g., interest in observing others interact) and negative Shyness (e.g., social anxiousness or lack of social interest) that might be relevant (Lane & Bowman, 2021). For instance, positive Shyness is positively associated with ToM, and negative Shyness is negatively associated with ToM (Lane & Bowman, 2021). There is a possibility that the Swedish translation of the items in the temperament questionnaire EAS/EASI captures anxious aspects of Shyness to a greater degree than other languages. However, the items do not suggest any such difference at face value. Nonetheless, Shyness is still interesting when it comes to understanding factors related ToM development, and further research is needed.

Activity

The lack of significant findings between Activity and ToM does not fit the general conclusion from Lane and Bowman (2021). However, there are previous findings that align with the null results. An example comes from Wellman, Lane et al. (2011b), where they investigated Rothbart's theory of temperament using the CBQ (and the Child behavior checklist) in relation to ToM in ages around 3–8 years of age. Even if they included a measure of Activity through the CBQ, no significant ToM relation surfaced. Another example is Mink et al. (2014), who (with extreme values in the data) initially found a significant negative association between Activity at 18 months and ToM at three years of age. However, after removing four extreme values, they found the relation no longer significant and concluded that "…the relation between Activity Level and ToM is not a meaningful one…" (p. 73). Nonetheless, further investigations of the relations between Activity and ToM are warranted.

The Lack of Stability of Temperament

The stability of temperament is important to consider when trying to understand the measures of temperament at young ages. It has been suggested that the older the child gets, the easier it may become to estimate temperament or rate traits accurately (Bould et al., 2013). For instance, Roberts and DelVecchio (2000) presented metanalytic results from 152 longitudinal studies that show a lack of

stability of temperament early in life, with consistency rising (but not consistently) per decade. The current data does support this line of reasoning. Notably, there are many high correlations within and between temperament dimensions and across measurements. However, there are still many instances of participants being relatively low in Emotionality, Activity, and Shyness at two years of age but then switching to being relatively high at four years of age (or vice versa). Because the temperament dimensions fluctuate for many of the included participants and ages, the results are compatible with findings showing that temperament stability might emerge later in life (Roberts & DelVecchio, 2000).

Social Factors in Relation to ToM

Analyses performed in *Study III* revealed that MST is related to ToM. Specifically, parental use of cognition words was the MST category most often related to ToM, but emotion and desire words were also associated with ToM. Similarly, all quantitative types of MST (i.e., absolute frequency, proportions, and vocabulary size) had instances of significant relations to ToM.

Parental MST in Relation to Children's ToM

In *Study III*, two novel findings within the field of ToM research were reported, primarily because of the type of statistical analysis used. First, the rate of change in ToM development from 3–5 years of age was positively associated with the absolute frequency of parental cognitive words spoken at two years of age. Second, the size of the emotional vocabulary used by parents at three years of age was also positively associated with the rate of change in ToM development from 3–5 years of age. The cognition word finding aligns with, supports, and extends previous results showing that parental use of cognition words, especially before age 3, is associated with FB (for a meta-analysis, see Tompkins et al., 2018). However, the emotion vocabulary finding contradicts the finding that parental use of emotion words is generally not associated with ToM development in the ages 0–5 (as highlighted by Tompkins et al., 2018).

Moreover, it is essential to emphasize that this is the first time MST vocabulary size has been investigated concerning ToM development. In *Study III*, when comparing emotional vocabulary with the absolute frequency of cognition words (which is highly related to ToM), the results suggest that parental emotion vocabulary size is similarly associated with ToM. This finding indicates that further and perhaps extended analyses of previous research might

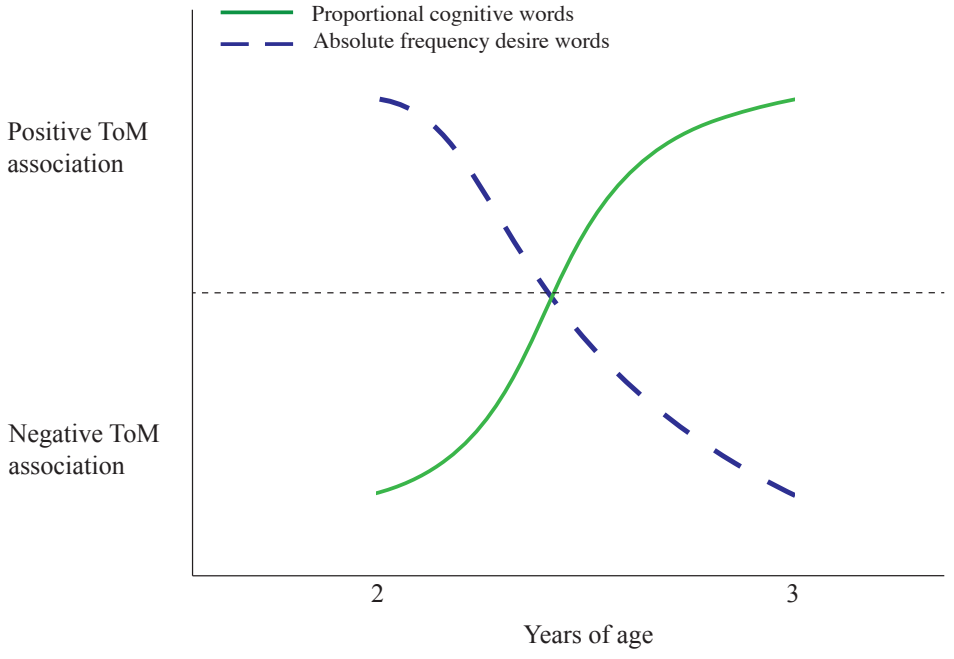
uncover another important aspect of parental MST that has passed unbeknownst to all since the start of investigations between MST and ToM. To iterate, in *Study III,* it is suggested that "…parents with larger emotion vocabularies might be better at describing the spectrum of relevant emotion states around the child. This suggestion, however, is a topic for future research and replication" (p. 20).

Other findings from *Study III* supporting or extending previous research were that proportions of cognitive words spoken by parents at three years of age were positively related to ToM ability at four years of age (for a meta-analysis, see Tompkins et al., 2018) and desire words spoken by parents at three years of age had a negative association with ToM at four years of age. In addition, negative associations between ToM and parental use of desire words have been reported previously (Chan et al., 2020; Taumoepeau et al., 2019; Taumoepeau & Ruffman, 2008). However, desire words are not always negatively associated with ToM, and relations between desire and ToM are not generally visible (Tompkins et al., 2018). Instead, the association between desire and ToM seems to be age-dependent, with the absolute frequency of parental desire words spoken positively related to ToM before age three and negatively associated with ToM development after age three (Taumoepeau & Ruffman, 2008).

Age Sensitivity in Cognition Words Spoken by Parents

A finding from *Study III* does not align with previous research. Specifically, parents' proportions of cognitive words spoken at two years of age were found to have a negative association with children's ToM at four years of age. This finding concludes that parents who talked proportionately more about cognitions had children with lower ToM. It seems as if no previous study reports a negative associations between parental use of cognitive MST words and ToM. This discrepancy can very well be related to the data handling of outliers (discussed below). Nonetheless, the interpretation of the negative associations between cognitive MST and ToM is that there might be benefits of focusing on parents talking about specific MST categories at certain ages. For example, if allowed to speculate and inversely mirror the findings about desire words, talking much about cognitions at two years of age and not giving enough context using non-MST words might result in too high a mental strain for the child (see Figure 7 below), while children might be more readily able to benefit from high proportions of parental cognitive words spoken at older ages.

Figure 7 - *Change in Association Between Parental MST and Children's ToM in Early Ages.*



Note. The Figure describes the associations between MST (the colored lines) and ToM (the y-axis) on a conceptual level between two and three years of age. The dashed horizontal line signifies the conceptual limit between positive and negative ToM associations for parental use of MST. The association between ToM development and proportional parental frequency of cognitive words spoken seems to be negative at two years of age but positive at three years of age. However, the association between absolute parental frequency of desire words spoken appears to be positive at three years of age but negative at two years of age.

Qualitative Aspects of Parental MST

Qualitative aspects of child-parent interaction that have not been included in the current thesis can be found in MST research. For instance, it is also common to analyze Mind-Mindedness (MM). MM is a related term that refers to the ability of a parent to ascribe correct mental states to a child as they are happening (Tompkins et al., 2018). In addition, a longstanding debate has not been settled on whether MST or MM is more associated with FB and ToM. However, Devine and Hughes (2019) found that only MST predicted FB, but both MM and MST had a weak concurrent relation to FB.

Other quality measures are clarifications, referent, and appropriateness. For instance, Tompkins et al. (2018) analyzed MST results that include clarifications (e.g., explaining why a person might be acting in a certain way). They report that clarifications were more related to FB than simple mental state mentions (e.g., saying, "He is thinking"). Additionally, referent (e.g., referring to the child's, parent's, or others' mental states) and appropriateness (e.g., correctly capturing and commenting on, commonly, the child's mental states) are associated with ToM (Chan et al., 2020; Meins et al., 2001; Symons et al., 2005; Tompkins et al., 2022). These findings suggest that MST can have qualitative importance for ToM (Tompkins et al., 2018).

There is no measure of the quality of the MST words mentioned in the current thesis. However, there was an attempt to capture this measure in the current data. That attempt concluded that there was a large overlap between the different MST categories in most statements. For instance, a parent could say, "I think the child might want to look like it is angry since it did not get an ice cream." In that one sentence, the parent mentioned a cognition word (think), an emotion word (angry), and a desire word (want) in the same sentence. Because of this overlap, the effort was abandoned, and the analyses presented in *Study III* were pursued instead.

Siblings and ToM

Study III includes an evaluation of the association between siblings and ToM. The results show that the number of siblings had no significant association at any measurement year. This finding goes against the general finding that siblings are important for more general social development and keeping friends (Downey et al., 2015; Downey & Condron, 2004) and FB ability (Devine & Hughes, 2018). The only explanation for this discrepancy is speculative at best. For instance, Downey and Condron (2004) clearly showed an association between having one

or two siblings and teacher measurements of the child's ability to (1) form and maintain friendships, (2) get along with people who are different, (3) comfort or help other children, (4) express feelings, ideas, and opinions in a positive way, and (5) show sensitivity to the feelings of others. Therefore, Downey and Condron (2004) are practically investigating emotional understanding, which can differ from the more cognitively tied ToM measured in the current thesis. Also, the ToM effect of having siblings that are close in age (i.e., child-aged siblings) is not present in the emotional understanding results reported by Downey and Condron (2004), further supporting the separation between ToM and emotional understanding.

SES and ToM

The results from *Study III* revealed a significant and close to systematic association with ToM. However, the association was limited to ToM ability at four years of age, but not to the rate of change in ToM development. Therefore, the current findings align with the meta-analysis by Devine and Hughes (2018), where SES was found to have a modest but significant association with FBU.

It has been shown that compound SES variables incorporated in analyses that do not include control of verbal ability have the strongest association with ToM development (Devine & Hughes, 2018). Additionally, investigations with children closer to seven years of age and studies that include a wider age range also had stronger findings connecting ToM with SES. Therefore, the lack of systematic findings between ToM and SES in the current thesis might be the result of (1) the included measure of SES (which just barely constitutes a compound variable), (2) the fact that productive language was included in the same analysis, (3) a relatively young sample, (4) and a relatively narrow age range between (and most certainly within) years of measurement.

Given the results presented and the remaining uncertainties, future studies are suggested to capture SES as a compound measure that includes education, income, and occupation, with options to include other related variables (Task Force on Socioeconomic Status, 2007). Additionally, it might be beneficial to have a wider age range of participants, preferably including older participants, as SES differences might become more pronounced with age.

Strengths and Limitations

The main strength of this thesis is that it is based on a longitudinal study that allows for control for between- and within-subject factors, unlike cross-sectional studies. Despite a relatively large longitudinal sample, the age variation at each measured age was kept relatively low compared to similar studies. However, some limitations should be addressed.

The Pandemic

Having to limit testing the last year due to the Covid-19 pandemic affected many aspects of the current thesis. Not being able to finish the last year meant that a complete third measurement of ToM could not be performed. Additionally, EF could not be measured a fourth time. However, the loss of EF measurement is not as detrimental to the project's quality as the loss of ToM data. This is because the fourth measurement of EF would be more of a proof of concept, and the ToM was envisioned to give new insight into the development of ToM in relation to other cognitive factors. However, all other measures were only minimally, or not, affected by the pandemic.

Scientific Considerations

The proper scientific procedure has become a focused discussion during the last two decades considering the replication crisis (Shrout & Rodgers, 2018). Related to the replication crisis are discussions of identifying researcher degrees of freedom (Simmons et al., 2011) and the widespread occurrence of questionable research practices (John et al., 2012). As a result, checklists have been published to exemplify, identify, and reduce the generally problematic scientific procedures many researchers use. The following sections will cover bullet points regarding researcher degrees of freedom and questionable research practices in relation to the current thesis.

Researcher Degrees of Freedom

Simmons et al. (2011) require five important statements from researchers for reviewers to identify questionable conduct more easily. First, what rule of terminating data collection was used, which is covered in all method sections of all studies and the thesis at hand? Second, they require 20 observations per condition in the case of a t-test. This is discussed in length under the heading *Power* below. Thirdly, all variables included in the study are exemplified in the

current thesis, but only the focused variables included in the current thesis are mentioned in the individual studies. Fourthly, no manipulations were performed in the included studies. Fifth, no observations were removed in *Study I*, and results with and without eliminated observations (e.g., outliers) are presented in *Study III* but not in *Study II*. Regarding excluded tests, it is clearly stated what variables were included in the other manuscripts being prepared or submitted at the submission of each manuscript (but not all tests measured in the thesis project). Sixth, only *Study I* can be considered to include analyses with covariates removed (i.e., with and without gender). The main reason is that many model fit indices were inappropriate for models that did not include covariates in *Study II* and *Study III*. However, *Study III* had an acceptable fit for the absolute frequency analysis when excluding covariates. In that analysis, the results were the same regarding the positively significant relationship between the absolute frequency of parental cognition words and the rate of change in the child's ToM development.

Questionable Research Practices

Simmons et al. (2011) presented a 10-point list measuring self-reported questionable research practices by more than 1400 American psychologists. The exhaustive list is (1) not reporting all dependent variables, (2) deciding whether to collect more data after checking if results are significant, (3) failing to report all study conditions, (4) stopping collection early because the desired result was found, (5) "rounding off" p-values to make them seem lower than the alfa level (e.g., .05) when it is higher (e.g., .053), (6) only reporting studies that "worked," (7) deciding whether to remove data in light of the results, (8) stating that an unexpected finding was anticipated, (9) claiming that results are unaffected by variables not included in the analysis, and (10) falsifying data.

Comments on all points above regarding the current studies and thesis are, in brief: (1) the ToM scale is the dependent variable, (2) testing at two years of age was stopped in July because very few new participants would be available for testing during the end of the summer, (3) was covered in the fourth point under the previous heading, (4) covered in the second point under the current heading, (5) each significance test was checked to the 5^{th} decimal and consistently reported correctly, (6) not applicable, but some tests were excluded because they did not "work" (e.g., Lazy Susan and CBQ), (7) analyses were presented without outliers because it was deemed most appropriate, (8) all unexpected findings are

stated as such, (9) not to the best of my knowledge, and (10) the current work includes no instance of falsified data.

Statistical Discussion

Scientific studies require that the analyses of the data are appropriately conducted. The following section elaborates on many relevant aspects of the analysis performed for the current thesis.

Control

A common way of identifying a confounding variable is to investigate the correlation the confounder has with the predictor included in the study (Carlson & Wu, 2012). However, it has been argued that this justification is flawed (Wysocki et al., 2022). Instead, the suggestion is that researchers should explicitly state how the variables are causally linked to justify including them as a confounding variable (e.g., Rohrer, 2018).

Regarding the current thesis, all studies have a clearly defined analysis structure. However, *how* the variables are *causally* linked is not well-defined. To help define the causal structure, Wysocki et al. (2022) describe (on pages 4–5) four variable types to explore when defining the causal structure that are also relevant to the current thesis, namely Confounder, Collider, Mediator, and Proxy. In general, the results of an analysis will be improved if the covariates included in the analysis are confounders. However, if included as a covariate, the other variable types contribute to different adverse effects on the analyses. A brief judgment of possible causal structures relevant to the *Study III* has been exemplified in brief (Table 6). In sum, there are possibly accurate model alternatives that were not analyzed. For instance, there is a chance that parental MST can bring about developments in EF and, perhaps more likely, language development (echoing research regarding Specific language impairment and deaf children). This mediator effect may consequently aid ToM development. There is also an ongoing debate regarding the directionality of influence between ToM and EF, with a lack of research perhaps obscuring the relation between early ToM and later EF (Wade et al., 2018). Recent research suggests that directionality may change with increasing age, with EF assisting ToM early and ToM assisting EF later in the early and middle childhood ages (Osterhaus et al., 2022).

Given the possible alternative models, it can be considered probable that appropriate covariates were predominantly included in the current thesis and that

appropriate models were analyzed and reported. The analyses also follow the necessary logic that confounders are measured earlier than or at the same time as the outcome. However, unobserved confounders, perhaps even confounders not yet identified, may still be affecting the results.

Table 6 - *Possible Causal Structures Between Predictor, Outcome, and Control Variables for Study III.*

X (MST) → Y (ToM)					
		C			
Control variable type	Causal structure	SES	Nr. older. Sib.	EF	Lang
Confounder	X ← C → Y	++	++	++	++
Collider*	X → C ← Y	–	–	+	–
Mediator	X → C → Y	–	–	+	++
Proxy	X → C Y	–	–	–	–

Note. In the table's first row, the confounder's causal structure is exemplified. The causal structure shows that the effect of the confounder affects both Mental state talk (MST) and Theory of Mind (ToM). All included control variables are assumed to be confounders in the analyses included in *Study III*. In the current example, SES, Number of older siblings, EF, and Language are all assumed to be valid and likely confounders of MST and ToM. X = Predictor; Y = Outcome; C = Control; SES = Socioeconomic status measured by averaged parental education.; Nr. older. Sib = Number of older siblings; EF = Executive function; Lang = Productive language; * = Applicable to the general field of research but not applicable to the longitudinal analyses performed in *Study III*, as no covariate included in the analysis was measured after the outcome variable.; – = A highly unlikely alternative causal structure; + = A valid alternative; ++ = A valid and likely alternative.

Missing Data

Full information maximum likelihood estimation (FIML) was utilized in *Study II* and *Study III* to handle missing data and outliers being removed. In addition, a more inclusive version of FIML (FIML.x) was chosen. The normal FIML removes incomplete cases before calculating the covariance matrix to approximate missing data. FIML.x, on the other hand, does not remove any cases and instead keeps all cases, thereby maximizing the information used in the approximation. As the number of complete cases was a bit lower than the total sample sizes, FIML.x seemed the most appropriate way to handle missing data in the current dataset. Nonetheless, when handling missing data and outliers, statistical methods generally implemented with more extensive datasets (N>200) were employed. This made analyses susceptible to non-optimal solutions that might not result in high-quality approximations of missing values (Rosseel, 2020).

Statistical Estimation

On a related note, the data in *Study II* and *Study III* were, even after outlier removal, found to be non-normally distributed. Therefore, robust estimators were used in the analyses in these studies. There are a few estimators, and a robust (Hubert-White) maximum likelihood estimator (MLR) was chosen. MLR performs well in samples with fewer than 200 participants that are not normally distributed (Li, 2016). This means that the current data fit both criteria. The measures and cut-offs used to ensure the current models have an appropriate fit are also common and recommended. Notably, none of the models had significant Chi^2-tests, and almost all had acceptable fit across all included fit indices, suggesting appropriate fit.

Power

Sufficient power is central to any planned or executed scientific study (Bakker et al., 2012). The required sample size to acquire the desired power is hinged on the effect size of the investigated relationship. Unfortunately, statistical power between factors can be diluted by factors such as non-normality, number of factors, and more complex models (Kline, 2016; Nicolaou & Masoner, 2013). Software is now readily available to calculate power in structural equation models (Jobst et al., 2021), and some rules of thumb have been published in relation to these diluting factors (Kyriazos, 2018). Two rules of thumb rely on

ratios between parameters included in the model and the number of participants (N). The first factor is the number of measured variables (p), and the other is the number of estimated parameters (q). Both suggest a factor 10:1 (i.e., 1 p or 1 q per 10 N) to be adequate to achieve necessary power (for a review, see Kyriazos, 2018), but N:q ratio might be best at, or higher than, 20:1 for latent variable models used on normal distributions and continuous outcomes (Jackson, 2003). However, recommendations for data similar to the data analyzed in the current thesis (i.e., MLR estimations with ordinal variables) are >200-500 (Bandalos, 2014).

When calculating the N:p and N:q ratio for the analyses that have been performed, *Study II* has a N:p of 8:1 (i.e., 121 participants to 14 measured variables) and a N:q of 4:1 (i.e., 121 participants to 14 measured variables and 28 estimated paths). For *Study III*, 80 participants were included, three ToM variables, six MST variables, and four time-invariant controls measured variables, giving 13 measured variables and an N:p ratio of 8:1. Additionally, *Study III* included 32 estimated parameters (in a 3-time LGCM with ten time-invariant controls, one observed variable at each time point, and two latent variables) giving a N:q ratio of 2.5:1.This suggests that the path model in *Study II* and LGCMs in *Study III* have low statistical power.

Statistical Conclusion

There is no way to ensure that the models included in the current thesis are, in fact, appropriate. As a result, there is a risk that some models are misspecified and, therefore, the conclusions are erroneous, especially as the sample sizes and statistical power in *Study II* and *Study III* are low (Rosseel, 2020). Still, as described in the previous paragraphs, removing outliers and thoughtfully implementing appropriate estimators reduced the risk of reporting misspecified models and erroneous results.

Reflections Regarding Tests

Including multiple tests that measure the same cognitive construct would have increased the chance of getting a stable measurement of that cognitive construct (e.g., Warnell & Redcay, 2019). Devine (2021) highlights that the ToM scale might not be optimally sensitive to individual differences, partly because of ceiling effects past the age of 5 and because of how it was designed (i.e., only measuring each step of the scale using one test). However, Wellman and Liu (2004) initially argued that investigations of individual differences using only

FB tests present a narrow representation of what ToM entails and that the ToM scale "… could provide a better measure to use in individual differences research examining the interplay between theory-of-mind understanding and other factors." (p. 524). Wellman and Liu's reasoning culminates in the conclusion that, in comparison to FB tests, "The current scale is usable with a wider range of ages, provides a more continuous variable for comparing individuals, and captures a greater variety of conceptual content. "(p. 537). Nonetheless, Karnell and Redcay (2019) suggest that "Future work should ideally include several items from a variety of scales…" (p. 7). Similarly, and more specifically, Devine (2021) suggests that the ToM scale would simply benefit from measuring each step "…using a range of task settings with different characters and materials" (p. 59). Therefore, a better way to capture ToM development could be to combine the strength of gradually measuring ToM using a scale with the increased ability to capture individual differences using repetitive measures of each scale step (including FB).

On a related note, many measurements of EF were included in the project (as premiered by Carlson, 2003) to increase the chance of getting a stable measurement of EF. Unfortunately, Lazy Susan (used widely as a WM measure) had lower and seemingly unreliable test-retest reliability than DCCS. Another task measuring EF was also implemented at two and three years of age. The task is called Baby Stroop (Hughes & Ensor, 2005) and is meant to be a useful test for very young samples. The task requires that the child is proficient at the game of "topsy-turvey." For example, if the experimenter asks the child to point to the (small) baby spoon, the child should point to the (big) mommy/daddy spoon. This is sequentially tested using baby and mommy/daddy cups. Unfortunately, the Baby Stroop task was noticeably more challenging than DCCS, where only 6 participants successfully completed the task at two years of age and only 44 participants at three years of age.

The reliability of EF measurements has been suggested to be affected by attentional issues, and combining EF measures to construct a single score of EF may be worse than keeping the tasks separate (Blair, 2016). However, given good reason to doubt the longitudinal stability of the Lazy Susan task (see Appendix II for a detailed explanation), it was not included in any analysis because of the risk of introducing unnecessary random variation to interpreting the relation between EF measures and ToM ability.

The current thesis would particularly have benefitted from multiple measurement methods being used to capture ToM development. As mentioned,

a nonverbal eye-tracking test of ToM ability (similar to Surian & Geraci, 2012) was included. Unfortunately, the test had random test-retest reliability. Unsurprisingly, the nonverbal eye-tracking test did not correlate with the verbal ToM measure (i.e., the ToM scale) as even ToM tests designed to capture ToM ability equally can give high ToM in one test, and low ToM on another for the same individual (Warnell & Redcay, 2019). Replication issues with similar eye-tracking tasks have also been highlighted (Boeg Thomsen et al., 2021; Dörrenberg et al., 2018; Kaltefleiter et al., 2021a; Kampis et al., 2021; Kulke et al., 2018). Kaltefleiter, et al. (2021a) investigated longitudinal trajectories of implicit ToM but failed to supply a graph of the individual trajectories in their study. Hence, only the developmentally and test-retest stable ToM scale was included as a ToM measure.

One MST-related realization might be relevant. My experience of testing parental MST is that MST outliers could also be "forced into existence" by not enforcing a hard stop of the conversation at around ten minutes but instead letting them finish talking. This led to some dyads talking for a longer time (e.g., 22 minutes) than others (e.g., the study average of eight minutes), naturally leading to parents uttering more words (and often also MST words) than a parent that talked for a shorter time period. However, even when enforcing this hard stop, it does not take care of the dyads not finding the conversation stimulating enough to keep it going for ten minutes. As stated earlier, when analyzing proportions, many of the issues discussed above are less of or not an issue at all.

Generalizability

Large resources were needed to complete this thesis and the longitudinal project it illuminates. The inclusion of other measurements of similar, related, or unmeasured concepts or constructs would have intensified the resource load. However, including other variables in the analyses could have altered the results and conclusions. Therefore, the current thesis (and the results described within) should be interpreted in the context in which it was performed by acknowledging the predictor, outcome, and control variables that were (and were not) included in each respective study.

Furthermore, the sample was geographically restricted to a small part of Sweden. Additionally, the sample has a higher education than the average in Sweden. This might have made the result more homogenous than expected from a sample representing the population better. Western, educated, industrialized, rich, and democratic (or WEIRD) cultures may not apply to the world's many

cultures and traditions (Keller, 2018). Therefore, the results in the thesis may primarily generalize to families with highly educated parents in Western cultures and contexts.

Ethical Considerations

All studies included in the current thesis were approved by the Swedish Ethical Review Authority (Dnr: 429–16) and were performed in accordance with the declaration of Helsinki. Participants were recruited using a national registry (i.e., SPAR) whilst opting out from receiving social security numbers, only asking for the mother's name and address. All parents signed informed consent forms and returned them to us via regular mail before being included in the study. After being included, their address, phone number, and names were linked to a participant number in a separate document on a password-protected data server. The file itself was also password-protected for added security. Anonymized raw data were stored in a locked, fire-proof cabinet or on a password-protected research data storage server.

The project was designed to include the relevant tests required for conducting meaningful analyses that can advance the field. None of the included tests were harmful to the children or their parents. Additionally, considerable effort was put into designing the procedure to be a pleasant experience for both the parent and the child.

At the start of each testing session, the participants were given general information about the project's aim, and the parents often asked follow-up questions that were answered in a general manner. Details about the investigated factors were not disclosed, even in response to exhaustive inquiries. The reason for not disclosing details was that detailed information about why certain measures were included in the project could have biased the parents' and child's behavior, at the lab and at home, towards the expected outcomes.

During the testing sessions, the parent and child were asked if they would like to continue or if they wanted to abort the session. Additionally, all participants were provided the right to withdraw from the study at any given time.

Theoretical Implications

There may be important insights to be gained by discussing the results of a study in relation to theoretical frameworks. However, there are occasions where the theoretical frameworks are not specific enough to aid the interpretation of the results. For instance, Baker et al. (2016) investigated the trajectory of ToM development month-by-month. They suggested that a more stable increase in ToM ability is probable in the long-term but very unstable in the short-term. They discuss that these findings do not fit any theory particularly well since no theory clearly defines what the trajectory of ToM development should be. However, the findings included in the current thesis may well fit within most of the theories included in the introduction, with an advantage for the ST/TT hybrid theory.

Simulation Theory / Theory Theory, and ToM

The theoretical perspective for this thesis is the hybrid theory of ST/TT (Apperly, 2008; Asakura & Inui, 2016; Harris, 2009; Mitchell et al., 2009). This theory assumes that a child's simulation ability increases with experience but will still struggle to grasp more advanced ToM tasks until the theory-building component has developed enough to detach their perspective from others. In other words, the child starts by getting to know their mind and then uses that experience to simulate better what others might think, feel, or know. In parallel, the ability to construct (implicit) abstract theories about one's own and other minds develops slower and will, when developed enough, allow for a more advanced ToM ability. This reasoning aligns well with Devine's (2021) suggestion that ToM ability is a composition of the child's propensity (which can be interpreted as the simulation ability) and fluency (which can be interpreted as the theory-building ability) of ToM.

Looking at the results from *Study I*, the individual performance suggests that FB is amongst the more demanding tasks to complete for the individual child, fitting the ST/TT account well. Specifically, DD, DB, and KA are all easier than CFB in all years of measurement. The CFB seems more challenging to complete than KA but much more complex than DD and DB (which show close-to-perfect performance at two and three years of age, respectively).

When diving into the more robust and most recent support of the ST/TT theory (Asakura & Inui, 2016), individual performance in the current thesis does not support their findings, i.e., that performance on DB and KA predict performance on CFB. However, considered in reverse, if a child succeeded in

CFB in *Study I*, it was very likely that the child had completed DB and KA. More specifically, on 46 of the 53 occasions a child completed CFB, they completed DB and KA (and DD). Five of the remaining seven occasions were children that completed CFB, DD, and DB (but not KA). No child completed only DB, KA, and CFB (which can be considered the main finding by Asakura & Inui, 2016). Only one child completed CFB and completed only one other step (KA). No child completed only CFB.

The results from *Study II* do fit within the ST/TT theory. The reason is that the only factor to surface when previous ToM was controlled for was the temperament dimension Shyness. In keeping with the general Social tendencies hypothesis (Lane & Bowman, 2021), the child may influence their social context through their temperamental composition. Therefore, the child's social interaction experiences are factors in developing ToM, in line with ST/TT.

The findings in *Study III* are equally well placed in an ST/TT framework. In this case, it was the environmental factor, or social factor, of parental MST that was found to be related to ToM. Parental MST may give the child better or lesser benefits in developing ToM. More specifically, the MST provided by parents allows for variation in the experience of exemplification of mental states. Additionally, the association between MST and ToM is not always positive, suggesting that the experience gained, perhaps concerning the child's current ToM ability or age, might not always be aiding further development. Again, this supports ST/TT, as experiences gained are instrumental for simulation ability and theory building.

In sum, parents' ability to capture ToM-related events around their children and children's temperamental behavior may influence the social context that the child regularly experiences. This conclusion supports ST/TT. However, the ST/TT theory is only one of many theories. The following text will introduce and further discuss the current thesis results.

Expression and Emergence, and ToM

One alternative theory to describe states that ToM development is linked with EF development. A related discussion revolves around two interpretations: if EF is a facilitator in the *expression* or the *emergence* of ToM (for a review, see Moses & Tahiroglu, 2010). If EF is a facilitator to the expression of ToM (e.g., Carlson et al., 1998), then EF assists ToM development. With this view, a child might have a functional understanding of other perspectives (i.e., possesses a ToM). Still, the child's performance on a FB task is limited by the ability to

refrain from answering questions based on their own perspective (i.e., limited EF does not allow expression of ToM). On the other hand, if EF is a facilitator of the emergence of ToM (e.g., Moses, 2001), then EF allows ToM to exist. This view suggests that EF capacity must reach a certain level before a child can be aware of perspectives that differ from their own. Taken together, that means that if changes in EF demand on a ToM task affect performance, then the expression view is supported (as shown by Setoh et al., 2016). Support for the emergence view is found when ToM tasks with minimal EF load are still related to EF. Both theories have empirical support, but neither can capture findings from all previous studies, which opens up the possibility that the theories are not mutually exclusive (Moses & Tahiroglu, 2010).

Study II was designed to investigate the association between ToM and EF, and the results were (primarily) insignificant. Additionally, EF demand is not manipulated in the ToM scale, meaning that no support could be found for the expression of EF. This means that the findings reported in the current thesis support neither of the EF accounts of social-cognitive development.

Nativist-Modular Account of ToM

The second alternative theory is another classic account of social cognition, based on the (non-social) *nativist-modular* account by Fodor (1983; who never mentions ToM in the same sense as Premack and Woodruff did), which states that the ToM module is specifically designed for social processing (Leslie et al., 2005; Scholl & Leslie, 1999, 2001). This results in the ToM module being virtually separate from other cognitive abilities (Scholl & Leslie, 1999). Scholl and Leslie (1999) describe the ToM module as an innate and separate part of the cognitive architecture affected by environmental cues. They write that environmental interactions may influence the exact timing of development while still highlighting that the ToM ability will be similar across individuals and cultures. Therefore, the module can be considered a cradle for the complete ToM understanding that a person can achieve, but the module's limitations are the same worldwide (Saxe, 2006).

The nativist-modularity theory is criticized for being unable to explain the effects of culture (e.g., Gopnik & Wellman, 2012; Wellman, 2014) as the strongest modular positions do not seem to allow for cultural variations in ToM development (Scholl & Leslie, 1999). However, the authors of the theory are clear on the subject: Environmental effects can hinder or aid the development of all aspects of ToM (Scholl & Leslie, 1999). In other words, the nativist-

modularity theory by Scholl and Leslie (1999) suggests that we all start with the same capacity for ToM, and environmental factors affect when the development of all aspects of ToM is triggered. They call it the "early theory of mind" (p. 697). Therefore, the environment does not make us develop different ToMs; it affects its timing.

Study II and *Study III* combined suggest that MST and temperament were the most relevant factors concerning ToM ability. Additionally, temperament can be considered an individual factor closely related to social experiences (Lane & Bowman, 2021). Therefore, the current thesis supports the environmentally dependent timing of ToM development, which the nativist-modularity theory suggests.

Interestingly, Wang et al. (2019) applied a Bayesian modeling framework to previously published FB research studies that investigated differences in ToM performance across different EF loads. Their results show that inhibition is crucial in refraining from focusing on the true belief interpretation of events. The authors suggest that their findings support the nativist/modular theory. Their approach to investigating and explaining ToM development in preschool children is somewhat different from that of Asakura and Inui (2016). However, the Bayesian models created in both publications present exciting perspectives and insights in the strive for a better understanding of ToM development at young ages. More work utilizing previous findings might be a fruitful way of helping the field advance.

Implicit/Explicit Theory of ToM

Another example of an alternative theory is Heyes and Frith's (2014) *implicit/explicit* account of ToM. They suggest that ToM development depends on a dual system: one implicit system that is active from birth and one explicit system that develops later. The implicit system is efficient, automatic, and independent of EF. The explicit system is slow, deliberate, and dependent on EF. During the first year of life, the infant becomes increasingly interested in faces (Frank et al., 2009), and by observation alone, the non-verbal (or implicit) ToM ability develops (Heyes & Frith, 2014). The explicit side develops later with underlying support from implicit ToM in a more socio-cultural manner through social interactions (Heyes & Frith, 2014). Heyes and Frith (2014) suggest that developing ToM (or learning to "mind-read") is very similar to learning to read. Contrary to nativist/modular theory, the implicit/explicit theory does not assume any specialized inherited neurocognitive mechanism for the explicit theory of

mind, but rather that the building blocks necessary to develop the explicit theory of mind are innately present in all humans (i.e., the "start-up kit"; Heyes & Frith, 2014).

The temperament and MST relations with ToM found in *Study II* and *Study III* are equally relevant for the Implicit/Explicit theory compared to the nativist-modularity theory. For this theory, the explicit system may gather information from the child's social interactions with parents and peers (influenced by the child's social tendencies) to further develop ToM. However, ToM dependency on EF receives limited support in the current thesis. Therefore, the suggestion of an EF-dependent explicit system is not supported. Additionally, there has been an onslaught of critique against previous research claiming successful measurement of implicit ToM (e.g., Heyes, 2014; Kulke et al., 2018). Notably, the implicit/explicit theory is not in question, but the research designs used in studies that support the theory are (Burnside et al., 2018; Dörrenberg et al., 2018; Heyes, 2014).

Similarities Between Theories

The similarities between the mentioned theories have already been discussed to some degree, but the theories (close to) universal dependence of language, EF, social and environmental influence, and neural architecture should at least be mentioned.

Language ability relates to all theories mentioned, with a crucial difference for the implicit system in the implicit/explicit theory. The implicit system cannot be measured using verbal tests (because verbal responses make a test explicit). Specifically, the implicit system is assumed to operate independently of language-tied deliberation and has a limited association with language-based instruction (Heyes, 2014). However, as stated initially, the other theories (including the explicit system) are related to language development.

Emergence and expression accounts of ToM are related to EF. The nativist/modular theory relies on EF, as the central assumption is that the ability to ignore one's perspective with the help of EF underlies the ability to solve ToM tasks correctly. The same ability to refrain from the personal perspective is apparent in all theories outlined in the current thesis. Additionally, the mere ability to understand the complexity of a social scenario (e.g., a FB task) relies upon keeping information with complex relations in mind. Hence, no ToM account is independent of EF (except the implicit system in the implicit/explicit theory), as inhibition, working memory, and cognitive flexibility are used

ubiquitously in the most basic everyday tasks. Still, the current thesis found EF less relevant to ToM than temperament and MST.

There is also the universal factor of environmental relations that are handled differently by the different theoretical outlooks. The theories less related to environmental factors or experiences are the emergence and expression and the nativist/modularity accounts. However, it is not unaffected by the environment or social experiences, merely less than the theories mentioned: implicit/explicit and ST/TT.

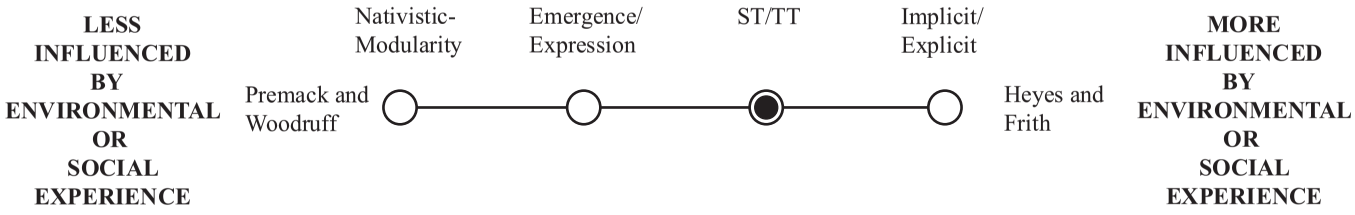
Finally, the current thesis did not present any new neuroscientific data, but all theories logically depend on some form of neural basis. Regarding explaining the architecture behind the theory, some theories (e.g., nativist-modularity theory) are more invested in explaining exactly how the architecture should be organized. Other researchers (who are not the original authors of the theories) occasionally need to introduce likely neural substrates (for a review, see Mahy et al., 2014). However, all theories benefit from describing clear neural definitions because this makes the theories more easily falsifiable and, in turn, better theories.

The Theoretical Conclusion

In sum, when considering the studies included in the current thesis and previous research on the ToM scale, it is difficult to arrive at a clear theoretical candidate that best encapsulates the variations in the development of ToM during the preschool years. Instead, all theories have merit, and often (at least when studying the original authors' description of their theories), no results that make the theories unfeasible. For instance, the current thesis has exemplified that dual-parallel theories (e.g., TT/ST, Implicit/Explicit, and, to an extent, the nativist-modular) are still viable.

What this thesis contributes to the theoretical discussion of ToM development is that the current results "shift the scale" towards a viable ToM theory leaning more towards social factors (or individual factors more related to social factors, such as temperament) and maybe less towards (other, more classical) individual factors (such as EF). Therefore, when considering the current and previous findings and comparing the outlined theories on a spectrum of social/environmental influence, with Premack and Woodruff's original ToM theory and Heyes and Frith's Theory as anchors, the ST/TT theory seems best able to capture the findings included in the current thesis (Figure 8).

Figure 8 - *A ToM Development Spectrum Including the Theories Outlined in the Current Thesis*



Note. The solid black dot signifies the theoretical position that best captures the findings included in the current thesis.

Gaps and Future Research

While working on the current thesis, a few interesting observations were made. The suggestions presented below are not always closely tied to the findings in the thesis, but they are still relevant to the discussion of which factors are related to ToM.

Gender Differences

A few studies have investigated gender differences using the ToM scale, but none have been longitudinal. All studies have been cross-sectional, and most have had small sample sizes. Six studies analyze and report no gender differences (Duh et al., 2016; Etel & Yagmurlu, 2015; Fujita et al., 2022; Rodrigues et al., 2015; Wellman et al., 2006; Wellman & Liu, 2004), and four have found gender differences (Calero et al., 2013; Hasni et al., 2017; Shahaeian, 2015; Sundqvist et al., 2018). It is unclear what the reason for the mixed findings could be. The sample in *Study I* has a high SES, but Shahaeian (2015) also included one sample with a high SES, and the current gender results are contradictory.

The main factor may be that studies have included different steps of the ToM scale. Given the current results in *Study I* (which included the four steps DD, DB, KA, and CFB), I suggest that the difference may stem from performance on the scale's HE or the EFB step. However, the difference in performance on HE or EFB might be specific to certain countries. For instance, Sundqvist et al. (2018) included a sample of Swedish children, Hasni et al. (2017) had a sample of children from the U.S., and Shaheian (2015) included a sample of Iranian children. All investigated a 5-step scale (DD, DB, KA, CF, and HE), and all studies found a gender difference. Still, Etel and Yagmurlu (2015) included a sample of Turkish children and, using the same scale steps, did not find any gender differences. Further research might illuminate possible gender differences in the specific ToM scale steps suggested above.

It is also possible that the differences between studies are related to the sample size and proportions of boys and girls included in the analyses.

Semantics and Syntax

Bringing the language and ToM discussion further, discussing what aspect of language might be most important for ToM development is important. As shown by de Mulder et al. (2019), it seems as if comprehension might be relevant.

Milligan et al. (2007), supported by Farrar et al. (2017), claimed that a more general measure of language (i.e., including comprehension, production, and syntax tests) has one of the strongest associations with ToM. However, one prominent theory of ToM development is that an intricate part of grammar might be at its center, namely complement clauses. A complement clause is a grammatic tool to convey, e.g., relationships between propositions and persons (e.g., "it's my toy"). Other clauses (e.g., "He thinks") can be added to these clauses to create more complex complement constructions (e.g., "He thinks [it's my toy]"). The important feature of these constructions is that they might be entirely correct, but the complement clause "it's my toy" might be incorrect (e.g., the toy might belong to someone else). The theoretical contribution by de Villiers (2007) reviewed past efforts and guided future ventures to evaluate the strength of association between different language measures. de Villiers (2007) argued (and got meta-analytical support from Milligan et al., 2007) that a child's ability to understand such complement clauses is essential to ToM performance.

Even if many studies support de Villiers's account of ToM development, recent studies have identified confounding design flaws in many supporting studies (for an extended discussion, see Boeg Thomsen et al., 2021; Fontana et al., 2018). Consequently, some researchers have attempted to investigate complement understanding in relation to ToM without confounding the results. For example, Kaltefleiter et al. (2021b) investigated syntax understanding with regard to FB in the months leading up to three years of age. The results were that complement ability was concurrently correlated with FB at 33 months of age, but previous complement ability did not correlate with FB ability at three years of age. Similarly, a longitudinal investigation by de Mulder et al. (2019) with Dutch children 2–3 years of age did not find relations between earlier complement ability and their later ToM battery scores. Still, they found that earlier comprehension (even if they called it general language) was related to later scores on their ToM battery.

In summary, few studies have found unambiguous support for relations between complement clauses and FB ability. The other studies either had confounded results or were performed with an unambiguous design and reported a lack of longitudinal relations. The main challenge for future intervention studies is to make sure they use an intervention design that can find support for the theory by de Villiers without confounding the results.

The issue with confounding results in complement clause research does not come to par with Dennet's (1978) criticism of Premack and Woodruff's (1978)

study, but that discussion is relevant. Whenever an effort is made to further our understanding of ToM, the investigation must be performed with a clear focus on protecting the validity of the ToM and FB ability being measured. If the validity is left unprotected even slightly, the interpretation of the results may well become problematic or invalid. As discussed above and in light of the methodological issues brought forth, Ruffman et al. (2003) might have summed it up in a most concise manner by stating, "…we found no evidence that syntax was more important than semantics" (p. 155).

ToM and EF, and Differences Between Countries

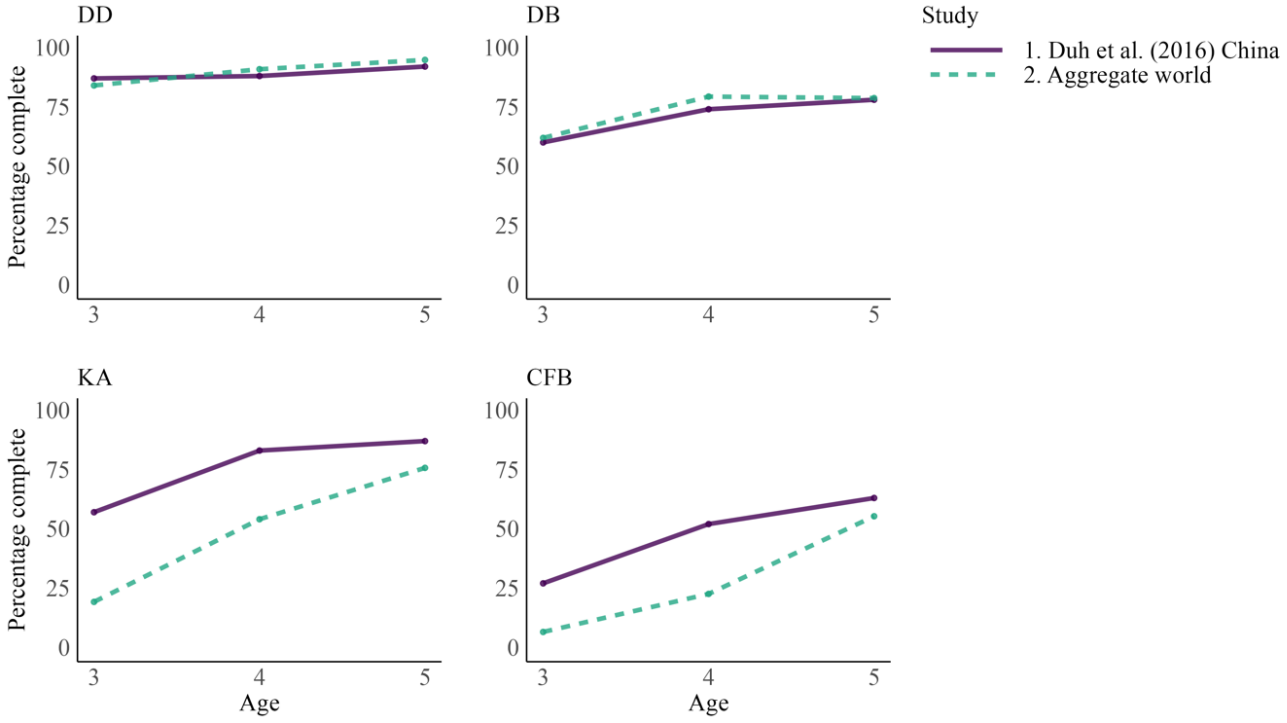
Compatible with the current findings, Fujita et al. (2022) found that their Japanese sample had a better EF ability but worse ToM understanding than a UK sample, carefully matched for age, verbal ability, gender, SES, and family structure. Fujita et al. (2022) state that their results challenge the EF theory's expression account (Moses, 2001). However, it is essential to highlight that Fujita et al.'s cross-sectional Japanese sample had an unusual developmental pattern when inspecting ToM scale performance on the DD and DB step. The analysis of differences between age groups is not presented in the study by Fujita et al. Still, I calculated Fisher's exact test for these comparisons by calculating the number of participants in each cell based on the presented performance percentages[6]. The pattern was that six-year-old children were numerically worse at these two tasks compared with three and four-year-old children, but it was only significant for DD compared between four- and six-year-old samples (Fischer, $p = .02$, Odds = 6.78, 95 % CI [1.08, 74.78]). This pattern is unexpected given how the increase in the performance, or clear ceiling effects, on these tasks is most often reported in previous studies for participants older than three years (e.g., Duh et al., 2016). Nonetheless, this oddity has affected Fujita et al.'s (2022) results in the direction that they report, that the Japanese children are worse at ToM tasks (but better at EF tasks) than their UK peers.

On a related note, something might have gone unnoticed in the ToM literature. When analyzing differences in FB performance between 2.5–7.5-year-old children, there is a comparable performance between mainland China and peers raised in the U.S. or Canada (Liu et al., 2008). However, when evaluating performance on individual steps in the ToM scale, Chinese preschoolers have a

[6] The actual data included a sampling zero and Fishers exact test can handle them (unlike Chi^2 tests) by increasing values in all cells by 0.5. The value of each cell was therefore increased by 0.5 before running the analysis.

clear advantage on the KA and CFB steps, especially at three and four years of age. This conclusion is based on the previous studies that can extract age-separated ToM scale results for children aged 3–5 years. The only difference in these steps is that KA and CFB performance is higher in China compared to many other countries when combined (see the aggregate comparison below in Figure 9). Differences may seem small in the Figure, but differences range from 11–37% for KA and 7–29 % for CFB at the individual ages. Differences for DD range from 2–3% and DB from 0–5%. Noteworthy is that Duh et al.'s study included 922 participants aged 3–5. All the other studies, together, included in the same graph, include 937 participants. The number of participants included in the individual age groups in Duh et al. and the aggregate studies in the graph are very similar. Specifically, Duh et al.'s study has 11 fewer, 19 fewer, and 15 more participants at three, four, and five years of age, respectively. In other words, by simply looking at the graph, it is possible to directly compare most of the available data investigating differences in development in four commonly used ToM scale steps. Therefore, it seems logical that well-founded conclusions regarding cultural differences in performance on the ToM scale can be drawn with that figure as support. Nonetheless, a more formal analysis of differences in ToM performance between countries might result in new insights into the relations between EF and ToM.

Figure 9 - *World Aggregate of ToM Scale Performance in Comparison to Duh et al. (2016)*



Note. Aggregate scores were calculated in three steps. First, I calculated the number of participants that successfully completed each step for each measurement year and each comparison study. Next, I calculated the sum of all participants that successfully completed each step for each year of measurement and each comparison study. Last, I calculated the percentage of completion for each step and year of measurement based on the sum of participants across all comparison studies (for each step and year of measurement). The comparison studies aggregated were Tomaya (2007), Qu & Shen (2013), Sundqvist et al. (2018), Henning et al. (2011), Fujita et al. (2022), and data from *Study I*.

The Underlying Outliers in MST Measurements

Tompkins et al. (2018) suggested that absolute frequency measures of MST "may be a more sensitive predictor of children's FBU compared to the proportion of MST." (p. 240), with FBU meaning FB understanding. The current findings do not quite support Tompkins et al.'s suggestion as only one significant result between absolute values and later ToM and multiple for proportions of MST was found. However, there is an important finding that should be mentioned. Analyses were performed with and without outliers in *Study III*. The data used for the manuscript was data without outliers. The supplementary files include a table with analyses of data with outliers still in the data. In the analyses with outliers included, significant associations between absolute values of MST but not with proportions (as is the common previously reported finding). Therefore, it seems uncommon for MST researchers to eliminate outliers from the data. The benefit of keeping outliers is that the analyzed data includes all available data points. However, as soon as the analysis is based on any statistical calculation, a few outliers can shift the results in favor of one conclusion exclusively present in the relation between outliers and the rest of the sample. Such a conclusion might be erroneous, as the relation between measurements in most samples could be different or even the opposite. As similar result to that previously reported was found using data with outliers and a different result using data with outliers removed, previous reports regarding MST and FB/ToM that might focus on the relations between masses of participants and odd outliers. However, there is a chance that absolute frequencies of parental MST are not the most "sensitive" but perhaps the most stable. Regardless of preschool age, it might be appropriate to speak to preschool children about cognition to widen and deepen their ToM understanding, unlike desires, whose effects dissipate after two years of age (as discussed above). Consequently, more research is needed to investigate the intricacies seemingly present in MST research.

The Possibility of a Globally Relevant Starting Pattern

One developmental sequence that may have been overlooked in the literature is the development of differences between the DB and KA steps at ages younger than four years. In line with the current findings, many previously published studies that included ToM scale measurements at two or three years of age (Duh et al., 2016; Henning et al., 2011; O'Reilly & Peterson, 2014; Qu et al., 2013; Tomaya, 2007; Wu & Su, 2014) show that children, on average, begin with a

"starting pattern" (i.e., DD>DB>KA>CFB), regardless of the country measured (e.g., China, Japan, Germany, Singapore) and later developmental divergence with regards to ToM scale pattern. There might only be two studies that deviate from the "starting pattern" at two or three years of age. The first (with three percentage points higher scores on DD than KA) is Sundqvist et al. (2018), and the second is a small sample (n = 8) of Aboriginal Australians in O'Reilly and Peterson (2014). For completeness, one study supporting the "starting pattern" hypothesis also has a relatively small sample size. Specifically, O'Reilly and Peterson's (2014) Anglo-Australian sample included 12 participants.

What would the theoretical implications of a "starting pattern" be? It seems as if it fits nicely into a nativistic-modular theory. Specifically, we all start with the same ToM, regardless of country of origin, to diverge based on environmental (i.e., cultural and social) cues. Nonetheless, and to reiterate, this finding can also support other theories. For instance, the implicit system might develop similarly for all children, and the influences of explicit ToM take over the implicit ToM ability at an older age (e.g., at 4). Equally possible, the theory component of ST/TT might "come online" at four years of age, making cultural and environmental effects more visible. In sum, the theoretical leaps within the preschool development of ToM remain elusive, even with the possibility of a global starting pattern of ToM development. Nonetheless, future studies are encouraged to investigate ToM scale performance in participants as young as two years of age to investigate the ubiquity of a ToM starting pattern on the group and perhaps even the individual level.

Observations and Insights

With all the information pertaining to the current thesis considered, the project offered a couple of insights that were more or less unexpected. Some may interest anyone who might design a similar project or study in the future. Hence, a list of observations, insights, and some possible solutions will be mentioned below.

The DCCS paradox

When testing 2-year-old children on the DCCS, many children failed the task. Some had a random response pattern during the task; some were very close to succeeding but perhaps lost focus on the last two trials. However, the most surprising participants were those who sorted cards (with the odd error) perfectly wrong. In other words, children who understood and completed the task in practice sorted according to the other dimension (shape) instead of the instructed dimension (color) during testing. Out of pure curiosity, I tested the eight children with this behavior as if they completed the first phase to see if the behavior persisted, which it did. The children with this behavior sorted almost consistently according to one dimension (shape) despite the first instruction they were provided and confirmed to understand was to sort according to another dimension (color). This observation can be understood using "attentional inertia" but maybe not in the classical sense (as described in Anderson, 1979), as the inertia I am describing here is somehow a result of individual information preference (or alike) being stronger than the instruction provided.

The Helpful Parent

One of the participants' parents was always sitting next to the child during testing. The parents were always instructed to "not help the child" during a test and to focus on interacting with the child in between tests. Also, the parents were assured that "the individual performance of your child is not our focus; it is the group's performance we are investigating." However, it became clear that parents generally want their children to succeed and are very supportive in the context that our project provided during each round of testing. It got to the point, at least for some of the families, that there was a need to pass instructions to the parent almost as often as to the child. This simultaneous and sometimes continuous instruction to the parent was needed for the data collected to measure the child's abilities and not how well the parent could assist the child in the event of minor adversities.

The solution to the situation described remains elusive since the behavior was present in the current project, at least for some families, even in the last year of testing. Nonetheless, the "issue" of the helpful parent is a reality.

The Unforgiving Nature of Longitudinal Studies

Depending on the research question, longitudinal projects may have enormous benefits over cross-sectional studies. However, in a longitudinal project, you are bound by your previous experiences and decisions. If you, perhaps in light of preliminary analyses, realize that one type of measure was overlooked and should have been included at an earlier measurement, there is seldom a way to collect that data post hoc. This type of "after the fact" knowledge can be demoralizing. However, the best way to handle these situations might be to spread the knowledge gained to others to avoid similar problematic hindsight. Hopefully, the current thesis has conveyed some of the issues that might arise in a longitudinal project involving preschool children and perhaps longitudinal research in general.

The Power of Utilizing Multiple Perspectives

While working on the current project, there appeared to be a seemingly endless source of knowledge from discussing the work with experts (and novices). Smaller or bigger parts have improved each time a new individual has been introduced to the project and provided constructive feedback. However, the more rewarding insights I attained while working on the current project did not come from discussing the information in the current thesis itself. Instead, they came from discussing issues and solutions in related research fields. Importantly, the benefit of learning from other's experiences is, quite logically, limited by the capacity to utilize them. Specifically, to see connections between others and your issues and solutions. Therefore, I want to highlight the importance of discussing minor and major research problems, regardless of which field they revolve in.

Conclusions

This thesis contributes to the understanding of social development in the preschool age. Specifically, there is a clear connection between earlier and later Theory of Mind ability, especially between immediately successive years of age. Theory of Mind development in young children is also associated with individual factors (such as executive functions, language development, and temperament) and social factors (such as socioeconomic status and parental ability to talk about others' knowledge, desires, and emotions). Still, when scrutinizing the results, it becomes clear that the individual factors associated with Theory of Mind often fade when the previous Theory of Mind ability is considered, while the social factors do not, especially when investigating the relations between parental mental state talk and the child's Theory of Mind development. Therefore, the current thesis shifts the scale towards Theory of Mind development being more related to social than individual factors. Nonetheless, it is important to state that the current findings are found in the current sample and with the included instruments and methods. More research is warranted regarding factors related to Theory of Mind development.

References

Aben, B., Stapert, S., & Blokland, A. (2012). About the distinction between working memory and short-term memory. *Frontiers in Psychology*, *3*(Aug), 1–9. https://doi.org/10.3389/fpsyg.2012.00301

Adrián, J. E., Clemente, R. A., & Villanueva, L. (2007). Mothers' use of cognitive state verbs in picture-book reading and the development of children's understanding of mind: A longitudinal study. *Child Development*, *78*(4), 1052–1067. https://doi.org/10.1111/j.1467-8624.2007.01052.x

Anderson, D. R. (1979). Active and passive processes in children's television viewing. *Paper Presented at the 87th Annual Meeting of the American Psychological Association*, 24.

Apperly, I. A. (2008). Beyond simulation-theory and theory-theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind." *Cognition*, *107*(1), 266–283. https://doi.org/10.1016/j.cognition.2007.07.019

Asakura, N., & Inui, T. (2016). A Bayesian framework for false belief reasoning in children: A rational integration of theory-theory and simulation theory. *Frontiers in Psychology*, *7*(Dec), 1–11. https://doi.org/10.3389/fpsyg.2016.02019

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, Alan. (1992). Working memory. *Science*, *225*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baker, S. T., Leslie, A. M., Gallistel, C. R., & Hood, B. M. (2016). Bayesian change-point analysis reveals developmental change in a classic theory of mind task. *Cognitive Psychology*, *91*, 124–149. https://doi.org/10.1016/j.cogpsych.2016.08.001

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

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 and Allied Disciplines, *42*(2), 241–251. https://doi.org/10.1017/S0021963001006643

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

Baron-Cohen, Simon, Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *38*(7), 813–822. https://doi.org/10.1111/j.1469-7610.1997.tb01599.x

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *The Handbook of Child Language* (pp. 95–151). Basil Blackwell. https://doi.org/10.1111/b.9780631203124.1996.00005.x

Bates, E., Hartung, J., Marchman, V., Thal, D., Fenson, L., Reilly, J., Dale, P., & Reznick, J. S. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123. https://doi.org/10.1017/S0305000900008680

Beaudoin, C., Leblanc, É., Gagner, C., & Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology*, *10*(January). https://doi.org/10.3389/fpsyg.2019.02905

Berglund, E., & Eriksson, M. (2000). Communicative development in Swedish children 16-28 months old: The Swedish early communicative development inventory - Words and sentences. *Scandinavian Journal of Psychology*, *41*(2), 133–144. https://doi.org/10.1111/1467-9450.00181

Biblarz, T. J., & Stacey, J. (2010). How does the gender of parents matter? *Journal of Marriage and Family*, *72*(1), 3–22. https://doi.org/10.1111/j.1741-3737.2009.00678.x

Blair, C. (2016). Developmental science and executive function. *Current Directions in Psychological Science*, *25*(1), 3–7. https://doi.org/10.1177/0963721415622634

Boeg Thomsen, D., Theakston, A., Kandemirci, B., & Brandt, S. (2021). Do complement clauses really support false-belief reasoning? A longitudinal study with English-speaking 2- to 3-year-olds. *Developmental Psychology*, *57*(8), 1210–1227. https://doi.org/10.1037/dev0001012

Borgström, K., von Koss Torkildsen, J., & Lindgren, M. (2015). Substantial

gains in word learning ability between 20 and 24 months: A longitudinal ERP study. *Brain and Language*, *149*, 33–45. https://doi.org/10.1016/j.bandl.2015.07.002

Bornstein, Marc H, & Hendricks, C. (2012). Basic language comprehension and production in over 100,000 young children from sixteen developing nations. *Journal of Child Language*, *39*(4), 899–918. https://doi.org/10.1017/S0305000911000407

Bould, H., Joinson, C., Sterne, J., & Araya, R. (2013). The emotionality activity sociability temperament survey: Factor analysis and temporal stability in a longitudinal cohort. *Personality and Individual Differences*, *54*(5), 628–633. https://doi.org/10.1016/j.paid.2012.11.010

Brink, K. A., Lane, J. D., & Wellman, H. M. (2015). Developmental pathways for social understanding: Linking social cognition to social contexts. *Frontiers in Psychology*, *6*(May), 1–11. https://doi.org/10.3389/fpsyg.2015.00719

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology*, *130*(11), 67–78. https://doi.org/10.1016/j.jecp.2014.09.010

Buckingham, J., Beaman, R., & Wheldall, K. (2014). Why poor children are more likely to become poor readers: The early years. *Educational Review*, *66*(4), 428–446. https://doi.org/10.1080/00131911.2013.795129

Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2018). Implicit false belief across the lifespan: Non-replication of an anticipatory looking task. *Cognitive Development*, *46*(February 2017), 4–11. https://doi.org/10.1016/j.cogdev.2017.08.006

Buschke, H. (1963). Retention in immediate memory estimated without retrieval. *Science*, *140*(3562), 56–57. https://doi.org/10.1126/science.140.3562.56

Buss, A. H., & Plomin, R. (1975). *A temperament theory of personality development*. Wiley-Interscience.

Buss, A. H., & Plomin, R. (1984). *Temperament: Early developing personality traits*. Erlbaum.

Calero, C. I., Salles, A., Semelman, M., & Sigman, M. (2013). Age and gender dependent development of theory of mind in 6- to 8-years old children. *Frontiers in Human Neuroscience*, *7*(May), 1–7. https://doi.org/10.3389/fnhum.2013.00281

Camaioni, L., Castelli, M. C., Longobardi, E., & Volterra, V. (1991). A parent report instrument for early language assessment. *First Language*, *11*(33), 345–358. https://doi.org/10.1177/014272379101103303

Carlson, K. D., & Wu, J. (2012). The illusion of statistical control: Control variable practice in management research. *Organizational Research Methods*, *15*(3), 413–435. https://doi.org/10.1177/1094428111428817

Carlson, S. M. (2003). Executive function in context: development, measurement, theory, and experience. *Monographs of the Society for Research in Child Development*, *68*(3), 138–51. https://doi.org/10.1111/j.1540-5834.2003.06803012.x

Carlson, S. M., Koenig, M. A., & Harms, M. B. (2013). Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(4), 391–402. https://doi.org/10.1002/wcs.1232

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., & 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*, *11*(2), 73–92. https://doi.org/10.1002/icd.298

Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development*, *69*(3), 672–691. https://doi.org/10.1111/j.1467-8624.1998.00672.x

Cassidy, K. W., Fineberg, D. S., Brown, K., & Perkins, A. (2005). Theory of mind may be contagious, but you don't catch it from your twin. *Child Development*, *76*(1), 97–106. https://doi.org/10.1111/j.1467-8624.2005.00832.x

Chan, M. H. ming, Wang, Z., Devine, R. T., & Hughes, C. (2020). Parental mental-state talk and false belief understanding in Hong Kong children. *Cognitive Development*, *55*(November 2019), 100926. https://doi.org/10.1016/j.cogdev.2020.100926

Chatham, C. H., Yerys, B. E., & Munakata, Y. (2012). Why won't you do what I want? The informative failures of children and models. *Cognitive Development*, *27*(4), 349–366. https://doi.org/10.1016/j.cogdev.2012.07.003

Colonnesi, C., Engelhard, I. M., & Bögels, S. M. (2010). Development in children's attribution of embarrassment and the relationship with theory of mind and shyness. *Cognition and Emotion*, *24*(3), 514–521. https://doi.org/10.1080/02699930902847151

Dale, P. S., Bates, E., Steven Reznick, J., & Morisset, C. (1989). The validity of a parent report instrument of child language at twenty months. *Journal of Child Language*, *16*(2), 239–249. https://doi.org/10.1017/S0305000900010394

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078. https://doi.org/10.1016/j.neuropsychologia.2006.02.006

De Mulder, H. N. M., Wijnen, F., & Coopmans, P. H. A. (2019). Interrelationships between theory of mind and language development: A longitudinal study of Dutch-speaking kindergartners. *Cognitive Development*, *51*(March), 67–82. https://doi.org/10.1016/j.cogdev.2019.03.006

de Villiers, J. (2007). The interface of language and theory of mind. *Lingua*, *117*(11), 1858–1878. https://doi.org/10.1016/j.lingua.2006.11.006

de Villiers, J. (2021). With language in mind. *Language Learning and Development*, *17*(2), 71–95. https://doi.org/10.1080/15475441.2020.1820338

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, *1*(4), 568–570. https://doi.org/10.1017/S0140525X00076664

Devine, R. T. (2021). Individual differences in theory of mind in middle childhood and adolescence. In R. T. Devine & S. Lecce (Eds.), *Theory of Mind in Middle Childhood and Adolescence: Integrating Multiple Perspectives* (pp. 55–76). Routledge. https://doi.org/10.4324/9780429326899-5

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development*, *85*(5), 1777–1794. https://doi.org/10.1111/cdev.12237

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., & Hughes, C. (2019). Let's talk: Parents' mental talk (not mind-

mindedness or mindreading capacity) predicts children's false belief understanding. *Child Development*, *90*(4), 1236–1253. https://doi.org/10.1111/cdev.12990

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.supp

Diamond, A. (2013). Executive functions. *The Annual Review of Psychology*, *64*(9), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

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*(February 2017), 12–30. https://doi.org/10.1016/j.cogdev.2018.01.001

Downey, D. B., & Condron, D. J. (2004). Playing well with others in kindergarten: The benefit of siblings at home. *Journal of Marriage and Family*, *66*(2), 333–350. https://doi.org/10.1111/j.1741-3737.2004.00024.x

Downey, D. B., Condron, D. J., & Yucel, D. (2015). Number of siblings and social skills revisited among american fifth graders. *Journal of Family Issues*, *36*(2), 273–296. https://doi.org/10.1177/0192513X13507569

Duh, S., Paik, J. H., Miller, P. H., Gluck, S. C., Li, H., & Himelfarb, I. (2016). Theory of mind and executive function in chinese preschool children. *Developmental Psychology*, *52*(4), 582–591. https://doi.org/10.1037/a0040068

Dunn, L. M., & Dunn, D. M. (2007). Peabody picture vocabulary test-Fourth edition. In *APA PsycTests*. https://doi.org/10.1037/t15144-000

Dunn, L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test-Revised.* American Guidance Service.

Ensminger, M. E., & Fothergill, K. (2003). A decade of measuring SES: What it tells us and where to go from here. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic status, parenting, and child development.* Lawrence Erlbaum.

Ensor, R., Devine, R. T., Marks, A., & Hughes, C. (2014). Mothers' cognitive references to 2-year-olds predict theory of mind at ages 6 and 10. *Child Development*, *85*(3), 1222–1235. https://doi.org/10.1111/cdev.12186

Ensor, R., & Hughes, C. (2008). Content or connectedness? Mother-child talk and early social understanding. *Child Development*, *79*(1), 201–216. https://doi.org/10.1111/j.1467-8624.2007.01120.x

Eriksson, M., & Berglund, E. (1999). Swedish early communicative development inventories: Words and gestures. *First Language*, *19*(55), 55–90. https://doi.org/10.1177/014272379901905503

Espy, K. A. (2004). Using developmental, cognitive, and neuroscience approaches to understand executive control in young children. *Developmental Neuropsychology*, *26*(1), 513–540. https://doi.org/10.1207/s15326942dn2601

Etel, E., & Yagmurlu, B. (2015). Social competence, theory of mind, and executive function in institution-reared Turkish children. *International Journal of Behavioral Development*, *39*(6), 519–529. https://doi.org/10.1177/0165025414556095

Farrar, M. J., Benigno, J. P., Tompkins, V., & Gage, N. A. (2017). Are there different pathways to explicit false belief understanding? General language and complementation in typical and atypical children. *Cognitive Development*, *43*, 49–66. https://doi.org/10.1016/j.cogdev.2017.02.005

Farrar, M. J., & Maag, L. (2002). Early language development and the emergence of a theory of mind. *First Language*, *22*(2), 197–213. https://doi.org/10.1177/014272370202206504

Fay-Stammbach, T., Hawes, D. J., & Meredith, P. (2014). Parenting influences on executive function in early childhood: A review. *Child Development Perspectives*, *8*(4), 258–264. https://doi.org/10.1111/cdep.12095

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), i. https://doi.org/10.2307/1166093

Fodor, J. A. (1983). *The modularity of the mind*. MIT Press. https://doi.org/10.7551/mitpress/4737.001.0001

Fontana, E., Adenzato, M., Penso, J. S., Enrici, I., & Ardito, R. B. (2018). On the relationship between theory of mind and syntax in clinical and non-clinical populations: State of the art and implications for research. *The Open Psychology Journal*, *11*(1), 95–104. https://doi.org/10.2174/1874350101811010095

Fu, I. N., Chen, K. L., Liu, M. R., Jiang, D. R., Hsieh, C. L., & Lee, S. C. (2023). A systematic review of measures of theory of mind for children. *Developmental Review*, *67*(1), 101061. https://doi.org/10.1016/j.dr.2022.101061

Fujita, N., Devine, R. T., & Hughes, C. (2022). Theory of mind and executive

function in early childhood: A cross-cultural investigation. *Cognitive Development*, *61*(September 2020), 101150. https://doi.org/10.1016/j.cogdev.2021.101150

Gagne, J. R., & Saudino, K. J. (2010). Wait for it! A twin study of inhibitory control in early childhood. *Behavior Genetics*, *40*(3), 327–337. https://doi.org/10.1007/s10519-009-9316-6

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

Goldsmith, H. H., Buss, A. H., Plomin, R., Rothbart, M. K., Thomas, A., Chess, S., Hinde, R. A., & McCall, R. B. (1987). Roundtable: What is temperament? Four approaches. *Child Development*, *58*(2), 505–529. https://doi.org/10.1111/j.1467-8624.1987.tb01398.x

Goldsmith, Harold Hill, & Rothbart, M. K. (1991). Contemporary instruments for assessing early temperament by questionnaire and in the laboratory. In A. Angleitner & J. Strelau (Eds.), *Explorations in Temperament: International perspectives on theory and measurement* (pp. 249–272). Plenum. https://doi.org/10.1007/978-1-4899-0643-4

Golombok, S. (2017). Parenting in new family forms. *Current Opinion in Psychology*, *15*, 76–80. https://doi.org/10.1016/j.copsyc.2017.02.004

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., Schulz, L., & Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. *COGSCI 2006: Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 1382–1387.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1–2), 145–171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Grape, A., & Sandstig, S. (2012). *Theory of mind, språkliga förmågor och ickeverbal intelligens hos barn mellan tre och fyra års ålder: Översättning och validering av theory of mind scale* [Master Thesis, Linköping University]. http://www.diva-portal.org/smash/get/diva2:540353/FULLTEXT01.pdf

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139–150. https://doi.org/10.2307/2086306

Hagekull, B., & Bohlin, G. (1990). *The Swedish translation of the EASI temperamental survey.* [Unpublished], Uppsala, Sweden: Uppsala

University.
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. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154. https://doi.org/10.1007/BF02172093
Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind & Language*, *7*(1–2), 120–144. https://doi.org/10.1111/j.1468-0017.1992.tb00201.x
Harris, Paul L. (2009). Simulation (mostly) rules: A commentary. *British Journal of Developmental Psychology*, *27*(3), 555–559. https://doi.org/10.1348/026151009X415484
Hasni, A. A., Adamson, L. B., Williamson, R. A., & Robins, D. L. (2017). Adding sound to theory of mind: Comparing children's development of mental-state understanding in the auditory and visual realms. *Journal of Experimental Child Psychology*, *164*, 239–249. https://doi.org/10.1016/j.jecp.2017.07.009
Henning, A., Spinath, F. M., & Aschersleben, G. (2011). The link between preschoolers' executive function and theory of mind and the role of epistemic states. *Journal of Experimental Child Psychology*, *108*(3), 513–531. https://doi.org/10.1016/j.jecp.2010.10.006
Hernandez, T. M., Aldridge, M. A., & Bower, T. G. R. (2000). Structural and experiential factors in newborns' preference for speech sounds. *Developmental Science*, *3*(1), 46–49. https://doi.org/10.1111/1467-7687.00098
Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science*, *17*(5), 647–659. https://doi.org/10.1111/desc.12148
Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091. https://doi.org/10.1126/science.1243091
Hiller, R. M., Weber, N., & Young, R. L. (2014). The validity and scalability of the theory of mind scale with toddlers and preschoolers. *Psychological Assessment*, *26*(4), 1388–1393. https://doi.org/10.1037/a0038320
Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002
Hoff, E., Laursen, B., & Tardif, T. (2002). Socioeconomic status and parenting. In Marc H. Bornstein (Ed.), *Handbook of Parenting* (2nd ed., Number

January, pp. 421–447). Erlbaum. https://doi.org/10.4324/9780429401459-13

Hollingshead, A. B. (1975). *Four factor index of social status.* [Unpublished], Yale University, New Haven, CT, USA.

Hou, X. H., Wang, L. J., Li, M., Qin, Q. Z., Li, Y., & Chen, B. Bin. (2022). The roles of sibling status and sibling relationship quality on theory of mind among Chinese preschool children. *Personality and Individual Differences*, *185*(September 2021), 111273. https://doi.org/10.1016/j.paid.2021.111273

Howard, A. A., Mayeux, L., & Naigles, L. R. (2008). Conversational correlates of children's acquisition of mental verbs and a theory of mind. *First Language*, *28*(4), 375–402. https://doi.org/10.1177/0142723708091044

Hughes, C. (1998). Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind? *Developmental Psychology*, *34*(6), 1326–1339. https://doi.org/10.1037/0012-1649.34.6.1326

Hughes, Claire, & Dunn, J. (1998). Understanding mind and emotion: Longitudinal associations with mental-state talk between young friends. *Developmental Psychology*, *34*(5), 1026–1037. https://doi.org/10.1037//0012-1649.34.5.1026

Hughes, Claire, & Ensor, R. (2005). Executive function and theory of mind in 2 year olds: A family affair? *Developmental Neuropsychology*, *28*(2), 645–668. https://doi.org/10.1207/s15326942dn2802

Hutchins, T. L., Prelock, P. A., & Bonazinga, L. (2012). Psychometric evaluation of the theory of mind inventory (ToMI): A study of typically developing children and children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *42*(3), 327–341. https://doi.org/10.1007/s10803-011-1244-7

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling*, *10*(1), 128–141. https://doi.org/10.1207/S15328007SEM1001_6

Jenkins, J. M., Turrell, S. L., Kogushi, Y., Louis, S., & Ross, H. S. (2003). A Longitudinal Investigation of the Dynamics of Mental State Talk in Families. *Child Development*, *74*(3), 905–920. https://doi.org/10.1111/1467-8624.00575

Jobst, L. J., Bader, M., & Moshagen, M. (2021). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods*. https://doi.org/10.1037/met0000423

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kaltefleiter, L. J., Schuwerk, T., Wiesmann, C. G., Kristen-Antonow, S., Jarvers, I., & Sodian, B. (2021). Evidence for goal- and mixed evidence for false belief-based action prediction in 2- to 4-year-old children: A large-scale longitudinal anticipatory looking replication study. *Developmental Science*, *August 2021*, 1–15. https://doi.org/10.1111/desc.13224

Kaltefleiter, L. J., Sodian, B., Kristen-Antonow, S., Grosse Wiesmann, C., & Schuwerk, T. (2021). Does syntax play a role in theory of mind development before the age of 3 years? *Infant Behavior and Development*, *64*(May), 101575. https://doi.org/10.1016/j.infbeh.2021.101575

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

Karlsson, E., & Östling, L. (2012). *Jämförelse mellan theory of mind- förmåga och pragmatisk förmåga hos svenska barn i 4 och 5 års ålder.* [Master Thesis, Linköping University]. http://liu.diva-portal.org/smash/get/diva2:530925/FULLTEXT01.pdf

Keller, H. (2018). Universality claim of attachment theory: Children's socioemotional development across cultures. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(45), 11414–11419. https://doi.org/10.1073/pnas.1720325115

Kirkham, N. Z., Cruess, L., & Diamond, A. (2003). Helping children apply their knowledge to their behavior on a dimension-switching task. *Developmental Science*, *6*(5), 449–467. https://doi.org/10.1111/1467-7687.00300

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (D. A. Kenny & T. D. Little (Eds.); 4th editio). The Guilford Press.

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

Koch, F. S., Sundqvist, A., Thornberg, U. B., Nyberg, S., Lum, J. A. G., Ullman, M. T., Barr, R., Rudner, M., & Heimann, M. (2020). Procedural

memory in infancy: Evidence from implicit sequence learning in an eye-tracking paradigm. *Journal of Experimental Child Psychology*, *191*. https://doi.org/10.1016/j.jecp.2019.104733

Kolk, M., & Andersson, G. (2020). Two decades of same-sex marriage in Sweden: A demographic account of developments in marriage, childbearing, and divorce. *Demography*, *57*(1), 147–169. https://doi.org/10.1007/s13524-019-00847-6

Korkman, M., Kirk, U., & Kemp, S. (1998). *NEPSY: A developmental neuropsychological assessment manual.* The Psychological Corporation.

Korucu, I., Selcuk, B., & Harma, M. (2017). Self-regulation: Relations with theory of mind and social behaviour. *Infant and Child Development*, *26*(3), 1–23. https://doi.org/10.1002/icd.1988

Kuczynski, L., Marshall, S., & Schell, K. (1997). Value socialization in a bidirectional context. *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory*, 23–50.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Kuhl, P. K. (2011). Early language learning and literacy: Neuroscience implications for education. *Mind, Brain, and Education*, *5*(3), 128–142. https://doi.org/10.1111/j.1751-228X.2011.01121.x

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, *29*(6), 888–900. https://doi.org/10.1177/0956797617747090

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, *09*(08), 2207–2230. https://doi.org/10.4236/psych.2018.98126

LaBounty, J., Bosse, L., Savicki, S., King, J., & Eisenstat, S. (2017). Relationship between social cognition and temperament in preschool-aged children. *Infant and Child Development*, *26*(2), 1–10. https://doi.org/10.1002/icd.1981

Lane, J. D., & Bowman, L. C. (2021). How children's social tendencies can shape their theory of mind development: Access and attention to social information. *Developmental Review*, *61*(July), 100977. https://doi.org/10.1016/j.dr.2021.100977

Lane, J. D., Wellman, H. M., Olson, S. L., Miller, A. L., Wang, L., & Tardif, T.

(2013). Relations between temperament and theory of mind development in the United States and China: Biological and behavioral correlates of preschoolers' false-belief understanding. *Developmental Psychology*, *49*(5), 825–836. https://doi.org/10.1037/a0028825

Lee, I. A., & Preacher, K. J. (2013). *Calculation for the test of the difference between two dependent variables with no variable in common [Computer software]*. http://quantpsy.org/corrtest/corrtest3.htm

Leppänen, P. H. T., Hämäläinen, J. A., Guttorm, T. K., Eklund, K. M., Salminen, H., Tanskanen, A., Torppa, M., Puolakanaho, A., Richardson, U., Pennala, R., & Lyytinen, H. (2011). Infant brain responses associated with reading-related skills before school and at school age. *Neurophysiologie Clinique/Clinical Neurophysiology*, *42*(1–2), 35–41. https://doi.org/10.1016/j.neucli.2011.08.005

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, *50*(1), 45–85. https://doi.org/10.1016/j.cogpsych.2004.06.002

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology*, *44*(2), 523–531. https://doi.org/10.1037/0012-1649.44.2.523

Longobardi, E., Spataro, P., D'Alessandro, M., & Cerutti, R. (2017). Temperament dimensions in preschool children: Links with cognitive and affective theory of mind. *Early Education and Development*, *28*(4), 377–395. https://doi.org/10.1080/10409289.2016.1238673

Longobardi, E., Spataro, P., Morelli, M., & Laghi, F. (2021). Executive function ratings in educational settings: Concurrent relations with cognitive and affective theory of mind. *Early Child Development and Care*, 1–13. https://doi.org/10.1080/03004430.2021.1975692

Madigan, S., Atkinson, L., Laurin, K., & Benoit, D. (2013). Attachment and internalizing behavior in early childhood: A meta-analysis. *Developmental Psychology*, *49*(4), 672–689. https://doi.org/10.1037/a0028793

Mahy, C. E. V., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience*, *9*, 68–81. https://doi.org/10.1016/j.dcn.2014.01.002

Manning, W. D., Fettro, M. N., & Lamidi, E. (2014). Child well-being in same-sex parent families: Review of research prepared for American Sociological Association amicus brief. *Population Research and Policy*

Review, *33*(4), 485–502. https://doi.org/10.1007/s11113-014-9329-6

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language*, *21*(2), 339–366. https://doi.org/10.1017/s0305000900009302

Marsiglio, W., Amato, P., Day, R. D., & Lamb, M. E. (2000). Scholarship on fatherhood in the 1990s and beyond. *Journal of Marriage and Family*, *62*(4), 1173–1191. https://doi.org/10.1111/j.1741-3737.2000.01173.x

Martin, R. M., & Green, J. A. (2005). The use of emotion explanations by mothers: Relation to preschoolers' gender and understanding of emotions. *Social Development*, *14*(2), 229–249. https://doi.org/10.1111/j.1467-9507.2005.00300.x

Mathiesen, K. S., & Tambs, K. (1999). The EAS temperament questionnaire - Factor structure, age trends, reliability, and stability in a Norwegian sample. *Journal of Child Psychology and Psychiatry*, *40*(3), 431–439. https://doi.org/10.1111/1469-7610.00460

McHale, J. P., & Rasmussen, J. L. (1998). Coparental and family group-level dynamics during infancy: Early family precursors of child and family functioning during preschool. *Development and Psychopathology*, *10*(1), 39–59. https://doi.org/10.1017/S0954579498001527

McHale, S. M., Updegraff, K. A., & Whiteman, S. D. (2012). Sibling relationships and influences in childhood and adolescence. *Journal of Marriage and Family*, *74*(5), 913–930. https://doi.org/10.1111/j.1741-3737.2012.01011.x

Meins, E., Fernyhough, C., Fradley, E., & Tuckey, M. (2001). Rethinking maternal sensitivity: Mothers' comments on infants' mental processes predict security of attachment at 12 months. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(5), 637–648. https://doi.org/10.1017/S0021963001007302

Meins, Elizabeth, Fernyhough, C., Arnott, B., Leekam, S. R., & De Rosnay, M. (2013). Mind-mindedness and theory of mind: Mediating roles of language and perspectival symbolic play. *Child Development*, *84*(5), 1777–1790. https://doi.org/10.1111/cdev.12061

Meins, Elizabeth, Fernyhough, C., Wainwright, R., Clark-Carter, D., Das Gupta, M., Fradley, E., & Tuckey, M. (2003). Pathways to understanding mind: Construct validity and predictive validity of maternal mind-mindedness. *Child Development*, *74*(4), 1194–1211. https://doi.org/10.1111/1467-8624.00601

Meltzoff, A. N. (1985). Immediate and deferred imitation in fourteen- and twenty-four-month-old infants. *Child Development*, *56*(1), 62. https://doi.org/10.2307/1130174

Meristo, M., Falkman, K. W., Hjelmquist, E., Tedoldi, M., Surian, L., & Siegal, M. (2007). Language access and theory of mind reasoning: Evidence from deaf children in bilingual and oralist environments. *Developmental Psychology*, *43*(5), 1156–1169. https://doi.org/10.1037/0012-1649.43.5.1156

Miller, C. A. (2006). Developmental relationships between language and theory of mind. *American Journal of Speech-Language Pathology*, *15*(May), 142–154. https://doi.org/10.1016/B0-08-044854-2/04198-5

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, S. A. (2016). Parenting and theory of mind. In *Parenting and Theory of Mind*. https://doi.org/10.1093/acprof:oso/9780190232689.001.0001

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Mink, D., Henning, A., & Aschersleben, G. (2014). Infant shy temperament predicts preschoolers theory of mind. *Infant Behavior and Development*, *37*(1), 66–75. https://doi.org/10.1016/j.infbeh.2013.12.001

Mitchell, P., Currie, G., & Ziegler, F. (2009). Two routes to perspective: Simulation and rule-use as approaches to mentalizing. *British Journal of Developmental Psychology*, *27*(3), 513–543. https://doi.org/10.1348/026151008X334737

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

Moeller, M. P., & Schick, B. (2006). Relations between maternal input and theory of mind understanding in deaf children. *Child Development*, *77*(3), 751–766. https://doi.org/10.1111/j.1467-8624.2006.00901.x

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H. L., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears,

M. R., Thomson, W. M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(7), 2693–2698. https://doi.org/10.1073/pnas.1010076108

Moriguchi, Y., Chevalier, N., & Zelazo, P. D. (2016). Editorial: Development of executive function during childhood. *Frontiers in Psychology*, *7*(Jan), 6–7. https://doi.org/10.3389/fpsyg.2016.00006

Moses, L. J. (2001). Executive accounts of theory-of-mind development. *Child Development*, *72*(3), 688–690. https://doi.org/10.1111/1467-8624.00306

Moses, L. J., & Tahiroglu, D. (2010). Clarifying the relation between executive function and children's theories of mind. In *Self- and Social-Regulation* (Vol. 15, Number 1, pp. 218–233). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195327694.003.0009

Nicolaou, A. I., & Masoner, M. M. (2013). Sample size requirements in structural equation models under standard conditions. *International Journal of Accounting Information Systems*, *14*(4), 256–274. https://doi.org/10.1016/j.accinf.2013.11.001

Nilsson, K. K., & de López, K. J. (2016). Theory of mind in children with specific language impairment: A systematic review and meta-analysis. *Child Development*, *87*(1), 143–153. https://doi.org/10.1111/cdev.12462

O'Reilly, J., & Peterson, C. C. (2014). Scaling theory of mind development in Indigenous- and Anglo-Australian toddlers and older children. *Journal of Cross-Cultural Psychology*, *45*(9), 1489–1501. https://doi.org/10.1177/0022022114542285

Osterhaus, C., & Bosacki, S. L. (2022). Looking for the lighthouse: A systematic review of advanced theory-of-mind tests beyond preschool. *Developmental Review*, *64*(May 2021), 101021. https://doi.org/10.1016/j.dr.2022.101021

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

Parke, R. D. (2004). Development in the family. *Annual Review of Psychology*, *55*(1), 365–399. https://doi.org/10.1146/annurev.psych.55.090902.141528

Pava, L. L. (2019). *The role of culture in theory of mind* [Doctoral Thesis. Edith Cowan University]. https://doi.org/10.1057/978-1-349-96042-2_425

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., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: You catch it from your sibs. *Child Development*, *65*(4), 1228–1238. https://doi.org/10.2307/1131316

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, *35*, 73–89. https://doi.org/10.1146/annurev-neuro-062111-150525

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

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, *4*(4), 515–526. https://doi.org/10.1016/j.celrep.2011.1011.1001.7.

Prime, H., Plamondon, A., & Jenkins, J. M. (2017). Birth order and preschool children's cooperative abilities: A within-family analysis. *British Journal of Developmental Psychology*, *35*(3), 392–405. https://doi.org/10.1111/bjdp.12180

Prime, H., Plamondon, A., Pauker, S., Perlman, M., & Jenkins, J. M. (2016). Sibling cognitive sensitivity as a moderator of the relationship between sibship size and children's theory of mind: A longitudinal analysis. *Cognitive Development*, *39*, 93–102. https://doi.org/10.1016/j.cogdev.2016.03.005

Qu, L., Shen, P., & Qianqian, F. (2013). Development of theory of mind in English-speaking Chinese Singaporean preschoolers. *Steering the Cultural Dynamics: Selected Papers from the 2010 Congress of the International Association for Cross-Cultural Psychology*, 85–94. https://scholarworks.gvsu.edu/iaccp_papers/107/

Ribner, A., Devine, R. T., Blair, C., & Hughes, C. (2022). Mothers' and fathers' executive function both predict emergent executive function in toddlerhood. *Developmental Science*, *July 2021*. https://doi.org/10.1111/desc.13263

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of

personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*(1), 3–25. https://doi.org/10.1037/0033-2909.126.1.3

Rodrigues, M. C., Pelisson, M. C. C., Silveira, F. F., Ribeiro, N. N., & da Silva, R. de L. M. (2015). Evaluation of theory of mind: A study with students from public and private schools. *Estudos de Psicologia (Campinas)*, *32*(2), 213–220. https://doi.org/10.1590/0103-166X2015000200006

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Röska-Hardy, L. (2009). Theory theory (simulation theory, theory of mind). In M. D. Binder, N. Hirokawa, & U. Windhorst (Eds.), *Encyclopedia of Neuroscience* (pp. 4064–4067). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-29678-2_5984

Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In R. van de Schoot & M. Miočević (Eds.), *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners* (1st ed., pp. 227–238). Routledge. https://doi.org/10.4324/9780429273872

Rothbart, M. K., & Bates, J. E. (2007). Temperament. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of Child Psychology, Sixth Edition: Social, Emotinal, and Personality Development.* (Vol. 3). Wiley. https://doi.org/10.1002/9780470147658.chpsy0303

Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review*, *34*(3), 265–293. https://doi.org/10.1016/j.dr.2014.04.001

Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers? Mental state language and theory-of-mind understanding. *Child Development*, *73*(3), 734–751. https://doi.org/10.1111/1467-8624.00435

Ruffman, T., Slade, L., Rowlandson, K., Rumsey, C., & Garnham, A. (2003). How language relates to belief, desire, and emotion understanding. *Cognitive Development*, *18*(2), 139–158. https://doi.org/10.1016/S0885-2014(03)00002-9

Sanson, A., Hemphill, S. A., & Smart, D. (2002). Emotions and social development in childhood. In P. K. Smith & C. H. Hart (Eds.), *Blackwell Handbook of Childhood Social Development: Second Edition* (pp. 97–

116). Blackwell Publishing. https://doi.org/10.1002/9781444390933.ch22

Sanson, A., & Rothbart, M. K. (1995). *Child temperament and parenting*. *January*.

Saxe, R. (2006). Why and how to study theory of mind with fMRI. *Brain Research*, *1079*(1), 57–65. https://doi.org/10.1016/j.brainres.2006.01.001

Schick, B., Villiers, P. De, Villiers, J. De, Hoffmeister, R., Development, S. C., Apr, M., Apr, N. M., Schick, B., Villiers, P. De, Villiers, J. De, & Hoffmeister, R. (2007). Language and theory of mind: A study of deaf children. *Society for Research in Child Development*, *78*(2), 376–396. https://doi.org/10.1111/j.1467-8624.2007.01004.x

Scholl, B. J., & Leslie, A. M. (1999). Modularity, development and "theory of mind." *Mind and Language*, *14*(1), 131–153. https://doi.org/10.1111/1468-0017.00106

Scholl, B. J., & Leslie, A. M. (2001). Minds, modules, and meta-analysis. *Child Development*, *72*(3), 696–701. https://doi.org/10.1111/1467-8624.00308

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293–327. https://doi.org/10.1037/bul0000303

Selcuk, B., Yavuz, H. M., Etel, E., Harma, M., & Ruffman, T. (2018). Executive function and theory of mind as predictors of socially withdrawn behavior in institutionalized children. *Social Development*, *27*(1), 109–124. https://doi.org/10.1111/sode.12252

Setoh, P., Scott, R. M., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(47), 13360–13365. https://doi.org/10.1073/pnas.1609203113

Shahaeian, A. (2015). Sibling, family, and social influences on children's theory of mind understanding: New evidence from diverse intracultural samples. *Journal of Cross-Cultural Psychology*, *46*(6), 805–820. https://doi.org/10.1177/0022022115583897

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive

psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Siriattakul, P., Suttiwan, P., Slaughter, V., & Peterson, C. C. (2021). Theory of mind (ToM) development in Thai deaf children. *Journal of Deaf Studies and Deaf Education*, *26*(2), 241–250. https://doi.org/10.1093/deafed/enaa036

Song, J. H., Volling, B. L., Lane, J. D., & Wellman, H. M. (2016). Aggression, sibling antagonism, and theory of mind during the first year of siblinghood: A developmental cascade model. *Child Development*, *87*(4), 1250–1263. https://doi.org/10.1111/cdev.12530

Statistics Sweden. (2020). Statistiska Centralbyrån (SCB). https://www.scb.se/en/finding-statistics/statistics-by-subject-area/education-and-research/education-of-the-population/educational-attainment-of-the-population/pong/statistical-news/educational-attainment-of-the-population-in-2020/

Sundqvist, A., Holmer, E., Koch, F. S., & Heimann, M. (2018). Developing theory of mind abilities in Swedish pre-schoolers. *Infant and Child Development*, *27*(4), 1–14. https://doi.org/10.1002/icd.2090

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

Suway, J. G., Degnan, K. A., Sussman, A. L., & Fox, N. A. (2012). The relations among theory of mind, behavioral inhibition, and peer interactions in early childhood. *Social Development*, *21*(2), 331–342. https://doi.org/10.1111/j.1467-9507.2011.00634.x

Symons, D. K., Fossum, K. L. M., & Collins, T. B. K. (2006). A longitudinal study of belief and desire state discourse during mother-child play and later false belief understanding. *Social Development*, *15*(4), 676–692. https://doi.org/10.1111/j.1467-9507.2006.00364.x

Symons, D. K., Peterson, C. C., Slaughter, V., Roche, J., & Doyle, E. (2005). Theory of mind and mental state discourse during book reading and story-telling tasks. *British Journal of Developmental Psychology*, *23*(1), 81–102. https://doi.org/10.1348/026151004X21080

Szpak, M., & Białecka-Pikul, M. (2019). Links between attachment and theory of mind in childhood: Meta-analytic review. *Social Development*,

November, 1–21. https://doi.org/10.1111/sode.12432

Tanaka, S. (2017). Intercorporeality and aida: Developing an interaction theory of social cognition. *Theory and Psychology*, *27*(3), 337–353. https://doi.org/10.1177/0959354317702543

Task Force on Socioeconomic Status, A. P. A. (2007). *Report of the APA task force on socioeconomic status*. American Psychological Association. http://www.ncbi.nlm.nih.gov/pubmed/21282485

Tasker, F. (2005). Lesbian mothers, gay fathers, and their children: A review. *Journal of Developmental and Behavioral Pediatrics*, *26*(3), 224–240. https://doi.org/10.1097/00004703-200506000-00012

Taumoepeau, M., & Ruffman, T. (2008). Stepping stones to others' minds: Maternal talk relates to child mental state language and emotion understanding at 15, 24, and 33 months. *Child Development*, *79*(2), 284–302. https://doi.org/10.1111/j.1467-8624.2007.01126.x

Taumoepeau, M., Sadeghi, S., & Nobilo, A. (2019). Cross-cultural differences in children's theory of mind in Iran and New Zealand: The role of caregiver mental state talk. *Cognitive Development*, *51*(May), 32–45. https://doi.org/10.1016/j.cogdev.2019.05.004

Teleman, U., Hellberg, S., & Andersson, E. (1999). *Svenska Akademiens grammatik* (First edit). Nordsteds.

Teti, D. (2002). Retrospect and prospect in the study of sibling relationships. In J. P. McHale & W. S. Grolnick (Eds.), *Retrospect and Prospect in the Psychological Study of Families* (pp. 193–224). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410604637

Thomas, A., Chess, S., Birch, H., & Hertzig, M. E. (1960). A longitudinal study of primary reaction patterns in children. *Comprehensive Psychiatry*, *1*(2), 103–112. https://doi.org/10.1016/S0010-440X(60)80014-4

Tomaya, K. (2007). Examining theory-of-mind tasks with Japanese children: The Wellman and Liu tasks. *Japanese Journal of Educational Psychology*, *55*, 359–369. https://doi.org/10.5926/jjep1953.55.3_359

Tompkins, V., Benigno, J. P., Kiger Lee, B., & Wright, B. M. (2018). The relation between parents' mental state talk and children's social understanding: A meta-analysis. *Social Development*, *27*(2), 223–246. https://doi.org/10.1111/sode.12280

Tompkins, V., Montgomery, D. E., & Blosser, M. K. (2022). Mother-child talk about mental states: The what, who, and how of conversations about the mind. *Social Development*, *31*(2), 281–302.

https://doi.org/10.1111/sode.12551

Van Bergen, P., & Salmon, K. (2010). The association between parent-child reminiscing and children's emotion knowledge. *New Zealand Journal of Psychology*, *39*(1), 51–56. https://www.researchgate.net/publication/281315779

Wade, M., Prime, H., Jenkins, J. M., Yeates, K. O., Williams, T., & Lee, K. (2018). On the relation between theory of mind and executive functioning: A developmental cognitive neuroscience perspective. *Psychonomic Bulletin and Review*, *25*(6), 2119–2140. https://doi.org/10.3758/s13423-018-1459-0

Walker, K. L., Ammaturo, D. A., & Wright, K. D. (2017). Are we assessing temperament appropriately? The emotionality activity sociability and impulsivity (EASI) temperament scale: A systematic psychometric review. *Canadian Psychology*, *58*(4), 316–332. https://doi.org/10.1037/cap0000108

Walker, S. (2005). Gender differences in the relationship between young children's peer-related social competence and individual differences in theory of mind. *Journal of Genetic Psychology*, *166*(3), 297–312. https://doi.org/10.3200/GNTP.166.3.297-312

Wang, L., Hemmer, P., & Leslie, A. M. (2019). A Bayesian framework for the development of belief-desire reasoning: Estimating inhibitory power. *Psychonomic Bulletin and Review*, *26*(1), 205–221. https://doi.org/10.3758/s13423-018-1507-9

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

Watson, A. C., Painter, K. M., & Bornstein, M. H. (2001). Longitudinal relations between 2-year-olds' language and 4-year-olds' theory of Mind. *Journal of Cognition and Development*, *2*(4), 449–457. https://doi.org/10.1207/S15327647JCD0204_5

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*(December 2020), 100945. https://doi.org/10.1016/j.dr.2020.100945

Wellman, H. M. (2002). Understanding the psychological world: Developing a theory of mind. In *Blackwell Handbook of Childhood Cognitive*

Development (pp. 167–187). Blackwell Publishers Ltd. https://doi.org/10.1002/9780470996652.ch8

Wellman, H. M. (2014). Making minds: How theory of mind develops. In *Oxford Scholarship Online*. https://doi.org/10.1093/acprof

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., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science*, *17*(12), 1075–1081. https://doi.org/10.1111/j.1467-9280.2006.01830.x

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

Wellman, H. M., Lane, J. D., LaBounty, J., & Olson, S. L. (2011). Observant, nonaggressive temperament predicts theory-of-mind development. *Developmental Science*, *14*(2), 319–326. https://doi.org/10.1111/j.1467-7687.2010.00977.x

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology*, *50*, 509–535. https://doi.org/10.1146/annurev.psych.50.1.509

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

Wolff, U. (2013). *Miniduvan: Test av fonologisk medvetenhet hos förskolebarn*. Hogrefe Psykologiförlaget.

Wu, Z., & Su, Y. (2014). How do preschoolers' sharing behaviors relate to their theory of mind understanding? *Journal of Experimental Child Psychology*, *120*(1), 73–86. https://doi.org/10.1016/j.jecp.2013.11.007

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, *5*(2). https://doi.org/10.1177/25152459221095823

Zelazo, P. D. (2006). The dimensional change card sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, *1*(1), 297–301.

https://doi.org/10.1038/nprot.2006.46

Zimmer-Gembeck, M. J., Webb, H. J., Pepping, C. A., Swan, K., Merlo, O., Skinner, E. A., Avdagic, E., & Dunbar, M. (2017). Review: Is parent child attachment a correlate of childrens emotion regulation and coping? *International Journal of Behavioral Development*, *41*(1), 74–93. https://doi.org/10.1177/0165025415618276

Appendices

Appendix I: Tests Not Included in the Current Thesis

The first three years of testing included tests not presented in this thesis. Unfortunately, Baby Stroop (Hughes & Ensor, 2005) had floor effects at two and three years of age. Episodic memory (Meltzoff, 1985) was also tested at three years of age, but with ceiling effects most likely due to testing being performed less than two hours after observation (in contrast to the original week-long interval). A language development test named DUVAN (Wolff, 2013) was included at three years of age as a pilot, and testing at four years of age led to a substantial number of incomplete attempts due to the task becoming very long if the child had a well-developed language skill. Testing three years of age also included a working memory test named "The farmhouse" (based on the Missing Scan task; Buschke, 1963), where a few animals are seen walking into a barn, and then all but one return. The task was for the child to say what animal was still in the barn. If they were successful, the number of animals that went into the barn increased, and they got two opportunities for each amount. At four and five years of age, another working memory test (Number repetition from NEPSY; Korkman et al., 1998). At four years of age, the Peabody Picture Vocabulary Test was administered (Dunn & Dunn, 2007, 1981). The strength of including the same measure of EF over time with DCCS and the ability to predict ToM performance at each measurement was prioritized over including more concurrent measures of EF. Partly because of analysis complexity but also because of the reliability associated with repeated measures using the same instrument. The serial reaction task (Koch et al., 2020) was used at four and five years of age and has been analyzed but does not fit the scope of the current thesis. A lengthy EEG procedure was performed two and three years of age but not included in this thesis at. The EEG task was designed to investigate neural signatures in response to differences in speech sounds and was administered similarly to Leppänen et al. (2011). The task was designed to capture the neuronal signals of a developing implicit sense of language perception. The collection of EEG data was difficult, with many children being unable to complete testing. These data are still being analyzed and could not be finalized for the thesis. Preliminary results are not in line with Leppänen et al. (2011).

Additionally, non-verbal ToM ability was measured at 2-, 3-, and four years of age using an eye-tracking task, similar to Surian and Geraci (2012). Analyses of the non-verbal ToM task became so vast that the analysis program (i.e., Tobii Studios) became unusable during testing at four years of age. The preliminary results from measurements at two and three years of age are in line with previous null findings (Boeg Thomsen et al., 2021; Dörrenberg et al., 2018; Kaltefleiter, Schuwerk, et al., 2021a; Kampis et al., 2021; Kulke et al., 2018). A summary of all tests included in the project for each year can be found in Appendix Table 1.

Appendix Table 1 - *Test Specification and Sample Size for all Tests and Years Measured.*

Instrument/ Method	Ability/Factor Measured	2 y. (2016)	3 y. (2017)	4 y. (2018)	5 y. (2019)	Test time (minutes)*
Hollingshead	SES - education	180	conf			3
Picture Book	Mental state talk	180	149			8
DI (pre)	Memory - Episodic					10
EEG (MMN)	Neural Lang signature					40
ToM&Jerry	ToM – Eye-tracking					20
SRT	Memory - Procedural					15
BabyStroop	EF - Inhibition					5
DCCS	EF - Shifting	138	149	134		7
LazySusan	EF - Working memory					10
The ToM scale	ToM - Verbal		142	134	53	10
Peabody	Lang - Comprehension					24
NEPSY	EF - Working memory					5
Duvan	Language - Phonology					25
The farmhouse	EF - Working memory					7
DI (post)	Memory - Episodic					3
SECDI	Language - Productive	164	130			35
EASI	Temperament	175		126		8
CBQ - VSF	Temperament					8

Note. The tests are ordered as they were presented to the participants. Each cell number shows the count of all data collected for that measurement and year.; y. = years old; DI = Delayed imitation; EEG = Electroencephalography; MMN = Mismatch negativity; SRT = Serial reaction task; DCCS = measure of Executive function; Lang = Language; ToM = Theory of Mind; NEPSY = Neuropsychological assessment; SECDI = Swedish Early Communicative Development Inventories; EASI = Emotionality, Activity, Sociability, Shyness, and Impulsivity Temperament Survey; CBQ-VSF = Child Behavior Questionnaire - Very Short Form; conf = participants SES was confirmed using follow-up questions; Blank = not tested; Gray = Tested, but not included in the current thesis; Green = tested; Yellow = tested but testing stopped due to Covid-19 pandemic.; * = the number is approximate as testing time varied between participants and the year of measurement.

Appendix II: Tests With Unexpected Issues

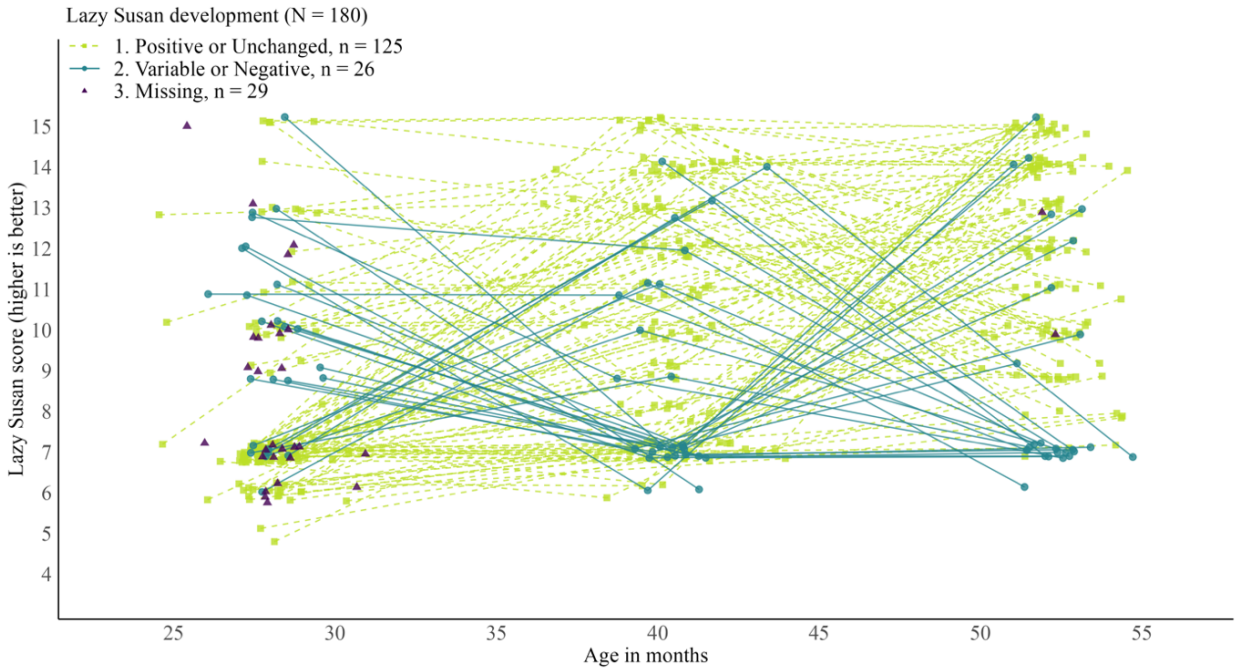
Inconsistent results were obtained with the Spin the Pots test (or the Lazy Suzan task; Hughes & Ensor, 2005), which is often used to measure WM (Appendix Figure 1). In the task, eight containers are placed on a rotatable tray and filled with one raisin each. Each cup had a lid, making it impossible to simply see if there was a raisin inside if standing up. The participant was supposed to find all the raisins. All cups were covered with a piece of cloth and rotated a random number of degrees, but always more than 360 degrees, between each trial. This task was administered when the children were two, three, and four years of age (i.e., all but the last measurement).

The intention was to include at least two tests that measure executive functioning to capture development better. However, when comparing performance on the Lazy Susan test to performance on DCCS, it became obvious that the performance on the Lazy Susan test was unreliable. A stability measure (see Lee & Preacher, 2013) was computed to compare the EF tests' ability to capture development. The Lazy Susan task was significantly worse in comparison. The analyses showed a steady and significant score increase between measurements between all measured years. However, a child could pass the test almost perfectly (e.g., a score of 14) one year and fail it the year after (e.g., a score of 7). Such instances were common in the data collected using the Lazy Susan task and almost non-existent in the DCCS (i.e., only seven participants performed worse on DCCS at a later year of measurement; see Appendix Figure 2). Inspecting the differences in raw scores for the Lazy Susan task, 43 participants had at least one lower score than the previous year's score. Speculatively, the reason behind the current findings might be that the Lazy Susan task captured some attentional process (e.g., the attentional process of long-term/phasic alerting; Petersen & Posner, 2012) to a higher degree compared to working memory development in the current sample. Therefore, the Lazy Susan test was not included in reporting the results.

Additionally, the current project included a temperament measure based on Rothbart's theory (Rothbart & Bates, 2007; Sanson & Rothbart, 1995), namely the Child Behavior Questionnaire - Very Short Form at four and five years of age. However, after analyzing the factor structure of the questionnaire at four years of age using confirmatory factor analyses and subsequent exploratory factor analyses, the additional temperament dimensions were deemed inconsistent for the current sample. Therefore, the questionnaire was dropped

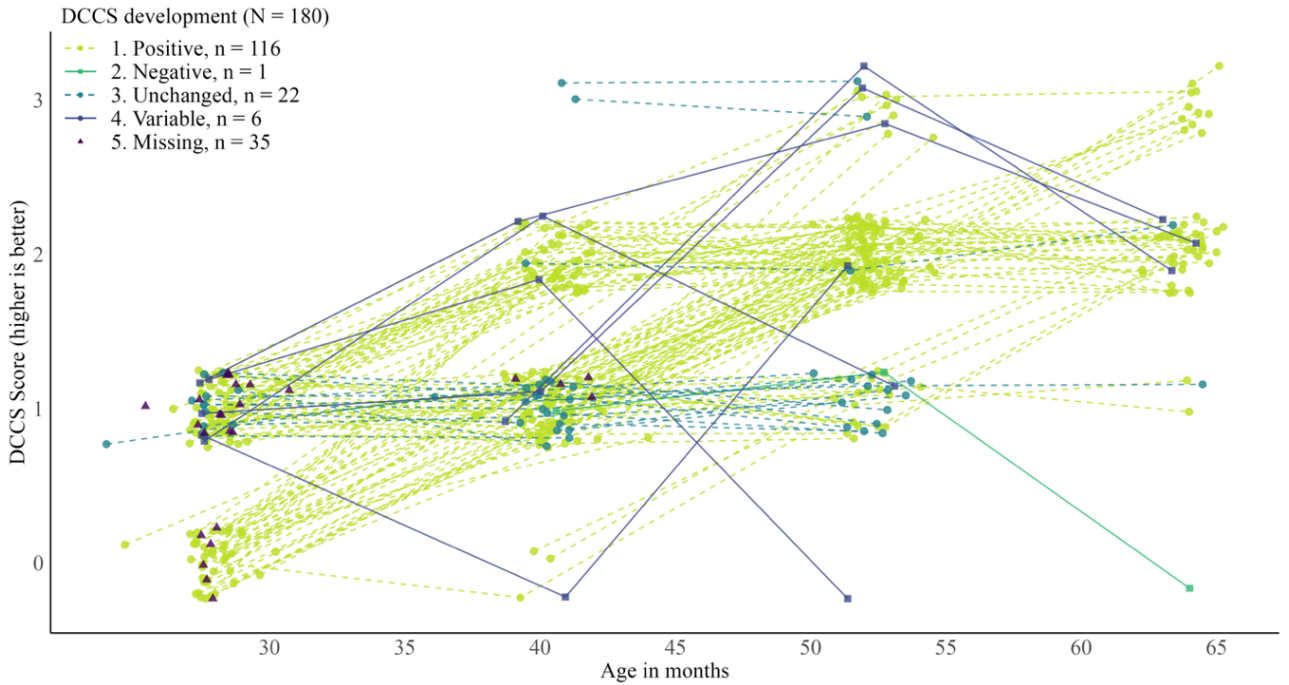
from *Study II* and is not included in any of the results reported in *Study II* in the current thesis.

Appendix Figure 1 - *Individual Performance on the Lazy Susan Task.*



Note. Missing = Participants that performed the Lazy Susan task once. The markers and their lines have been moved slightly on the y- and x-axis for easier interpretation. All solid lines show participants that succeeded (i.e., gained a score of eight or higher) on the task one year before but failed the task (i.e., gained a score of lower than 8) a later year. Of the 180 participants who participated at two years of age, 26 (15%) were such participants. A few participants perform almost perfectly at two years of age, fail at three years of age, and perform almost perfectly at four years of age, and vice versa. Additionally, 43 additional participants performed numerically worse than they did on a previous measurement at some age.

Appendix Figure 2 - *Individual Performance on the Comparison Task, the Dimensional Change Card Sort Task (DCCS).*



Note. Missing = Participants that performed the DCCS task once. The markers and their lines have been moved slightly on the y- and x-axis for easier interpretation. All solid lines show participants who succeeded in the task one year but failed a later year. There were seven such participants. Additionally, the remaining participants do not perform worse at any age.