Methods & Measures

ijbd

International Journal of Behavioral Development
2022, Vol. 46(3) 251–259
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01650254221077334
journals.sagepub.com/home/jbd
SAGE

Advanced scaling and modeling of children's theory of mind competencies: Longitudinal findings in 4- to 6-year-olds

Christopher Osterhaus[1,2], Susanne Kristen-Antonow[1], Daniela Kloo[1] and Beate Sodian[1]

Abstract

First-order theory of mind (ToM) development has shown to conform to a Guttman scale, with desire reasoning developing before belief reasoning. There have been attempts to test for internal consistency and scalability in advanced ToM, but not over a broad age range and only with a limited set of tasks. This 2-year longitudinal study ($N = 155$; $M_{age} = 4.2$; $SD = 0.85$ months; 68 girls, 87 boys) tests for the scalability of a broader range of ToM tasks, and we model the developmental *transition* from first-order to advanced ToM in 4- to 6-year-olds. Rasch analyses showed that psychometrically sound and developmentally sequenced scales emerge when measures of morally relevant and second-order false belief, as well as mental verb understanding, metacognition, and recognition of nonliteral speech are included. Individual differences were moderately stable over time, and there were systematic transitions from failure to success in children's performance, suggesting that conceptual continuity exists between first-order and advanced ToM.

Keywords

Extended theory of mind scale (extended ToM scale), advanced theory of mind (advanced ToM), Rasch analysis, transition from first-order to advanced ToM

Theory of mind (ToM) comprises diverse mindreading skills, including the ability to ascribe mental states to self and other (Wellman, 2020). Explicit ToM skills (as assessed with perhaps the most prominent measure of ToM, the false belief [FB] task) typically emerge around 4 years (Wellman & Liu, 2004). A large body of evidence suggests that there is a stable sequence in which children attain the conceptual insights that are involved in ToM: Initially, children develop the proficiency to reason about diverse desires (DDs); once they attain this skill, they come to understand that people may hold diverse beliefs (DBs), that knowledge access (KA) influences ignorance, that people sometimes hold FBs, and that we can hide our true emotions, that is, hidden emotions (HEs). This developmental pattern (DD > DB > KA > FB > HE) is reflected by the classic ToM scale (Wellman & Liu, 2004), for which Guttman and Rasch scaling have indicated a stable sequence across populations, including children from different cultures (e.g., Kristen et al., 2006; Shahaeian et al., 2011) and children with hearing impairment (Peterson et al., 2012).

Advanced ToM refers to more-complex mindreading skills that develop in middle childhood (e.g., Hughes & Devine, 2015). While some define it as the ability to put into use own ToM knowledge in fast, flexible, and accurate ways (e.g., Apperly, 2012), for many authors, advanced ToM implies changes at the conceptual level (e.g., Lagattuta et al., 2016). Similar to first-order ToM, there have been attempts to model these changes at the conceptual level, testing for the conceptual continuity, internal consistency, and scalability in advanced ToM (Hayward & Homer, 2017; Osterhaus et al., 2016; Peterson et al., 2012; Warnell & Redcay, 2019). These studies have, however, not been conducted over a broad age range (from first-order to advanced ToM) and only with a limited set of tasks. Peterson et al. (2012) have extended the classic ToM scale and introduced an additional step. This additional step involves children's understanding of sarcasm, that is, whether they understand that people can mean the opposite of what they say. This task is modeled after the strange stories (Happé, 1994), which assess children's understanding of nonliteral speech across a broader range of situations, such as in jokes, figures of speech, or lies. As expected, the sarcasm task proved to be more difficult than the other ToM tasks, and 41% of typically developing children aged 7.5–11.5 years solved it correctly. The number of steps taken by the children was significantly correlated with their age (Peterson & Wellman, 2019).

Other researchers have reported correlational evidence that can be interpreted as indicative of a low degree of conceptual coherence in first-order and advanced ToM: Warnell and Redcay

[1] Ludwig-Maximilians-Universität München, Germany
[2] University of Vechta, Germany

Corresponding author:
Christopher Osterhaus, Developmental Psychology in Education, Faculty of Education and Social Sciences, University of Vechta, Driverstr. 22, 49377 Vechta, Germany.
Email: christopher.osterhaus@uni-vechta.de

(2019) report no significant associations in 4- and 6-year-olds between an FB index (based on assessments of contents, explicit, and second-order FB) and children's understanding of HEs, as well as prominent measures of advanced ToM, including the eyes and faux-pas tests. These results are contrasted by findings by Osterhaus and Koerber (2021), who modeled advanced ToM competencies using factor analysis and who report, for children aged 5–8 years, significant and independent correlations between first-order and three different advanced ToM factors.

Correlational analyses alone, however, may not suffice to make the case for a conceptual continuity between first-order and advanced ToM—after all, the scale analyses of ToM tasks show that children's ToM development is sequential, that is, minimal coherence is to be expected when composite scores for tasks of diverse difficulties are correlated within narrow age or competence groups. To establish that there is conceptual continuity between ToM and advanced ToM, researchers need to ask (1) whether the classic ToM scale can be extended by adding age-appropriate, developmentally sequenced (i.e., more diverse and more difficult) tasks, (2) whether the ability indexes that are obtained from these developmentally sequenced scales produce estimates of individual differences that are longitudinally stable, and (3) whether there are systematic changes from failure to success (i.e., do children get better over time?) (Wellman & Liu, 2004).

The present study addresses these three questions. In particular, our 2-year longitudinal study involving 155 children tests for the scalability of a broader range of first-order ToM tasks than were included in the classic ToM scales. Following Wellman and Liu (2004), we use Rasch modeling to establish that the items in our extended ToM scale are of increasing difficulty. The unidimensional Rasch or one-parameter logistic (1-PL) model (Rasch, 1960) is a probabilistic model that assumes that all items in a test measure a single dimension of ability, and that the probability that test takers correctly solve a given item can be described by their ability and the difficulty of that item: If the former exceeds the latter, test takes will answer that item correctly (Kubinger, 2005). The Rasch model can therefore be regarded as a probabilistic analogue to Guttman scale analysis, which shares the notion that a person with a given ability level will likely (Rasch) or definitely (Guttman) only give correct responses to those items whose difficulty estimates are lower than that person's ability level (Wellman & Liu, 2004). Unlike the two-parameter logistic (2-PL) model (Birnbaum, 1968), the Rasch model assumes equal discriminations across items, rendering item difficulty the only item parameter postulated by the model. If the Rasch model holds, researchers can therefore conclude that there are no possibly unknown dimensions that affect item discrimination, and there is comparability of scores across tests. Also, under the fit of the Rasch model, relative performance differentials can be interpreted in a straightforward way, where equal trait-level distances are reflective of equal relative differences in performance, regardless of the item difficulty level (Embretson, 1996).

In addition to Rasch analysis, we attempt to model the developmental transition from first-order to advanced ToM in the age range from 4 to 6 years. To this end, we include additional first-order ToM tasks in the analysis, which allows to more reliably model individual differences in first-order ToM understanding, accounting for possible task-specific influences of performance, which may vary as a function of the specific context and target of the mental state ascription required by the task. Based on a definition of ToM as a conceptual framework encompassing mental state ascriptions to both self and others, we included measures of children's understanding of morally relevant FB (Killen et al., 2011) and of mental verbs of certainty (Moore et al., 1989; i.e., different contexts), as well as a measure of metacognition of own ignorance (i.e., Rohwer et al., 2012; different targets). Advanced ToM was assessed with measures of second-order FB understanding (Coull et al., 2006; Sullivan et al., 1994) and understanding of nonliteral speech (Happé, 1994). To test whether the stability in individual differences over time is independent of general information-processing skills (Carlson & Moses, 2001; Milligan et al., 2007), we also conducted assessments of verbal IQ and inhibition.

Methods

Participants

The participants were $N = 155$ (68 girls, 87 boys) healthy children who had participated in a longitudinal study of social-cognitive development from the age of 7 months. The children were recruited from public birth records, and they were from low to high middle-class families in a large city in southern Germany. Children's ages across the waves were as follows: During Wave 1, the children were 4 years old ($M = 50.1$ months, $SD = 0.86$); during Wave 2, they were 5 years old ($M = 60.1$ months, $SD = 0.65$); and during Wave 3, they were 6 years old ($M = 70.4$ months, $SD = 0.52$). Data collection was carried out between 2015 and 2017.

One quarter (25.4%) of the mothers had attended secondary school up to Grade 10 (not college-bound degree), 22.8% had attended secondary school up to Grade 12 (college-bound degree), and 51.8% had obtained a bachelor's degree or higher. Parents' informed consent was obtained from all participants. The present study was approved by the ethics committee of the Faculty of Psychology and Education, Ludwig Maximilian University, Munich, Germany (2013-11-11).

Procedure

Caretakers accompanied children to the assessments (conducted in German language) in a university laboratory.

Materials

An overview of the tasks that were used in Waves 1–3 is given in Table S-1 (in the online supplementary material).

ToM Scale. Children's first-order ToM was assessed with (a selection of tasks from) the ToM scale (Wellman & Liu, 2004). All six tasks were used at Wave 1; at Waves 2 and 3, the three most difficult ones were used (Tasks 4–6):

In the *DD* task (1), the children learn about a protagonist who has a preference contrary to their own. The children have to correctly predict what the protagonist would choose based on his preference (1 point). In the *DB* task (2), the children have to correctly predict the belief of another person that is contrary to their own (1 point), and the *KA* task (3) assesses their understanding that someone with full access to all relevant information will hold a different belief from someone without such KA (1 point).

In the *contents FB* task (4), the children are presented with a familiar Smarties box, which does not contain the popular sweets but a little toy pig. The children have to predict the FB of someone who has never looked inside the box (1 point). In the *location FB* task (5), the children are told that the mittens of a protagonist are in her backpack, but she believes they are in the closet. The children have to predict where she will look for her mittens, in the backpack or closet (1 point)? Finally, in the *HE* task (6), the children have to show that they understand that people may display emotions that differ from what they feel (1 point).

Test and control questions had to be answered correctly. Cohen's κ for interrater agreement ranged from .94 to perfect agreement across all ages.

Morally Relevant FB. In the authorized German version of the gender-matched accidental transgressor vignette (Killen et al., 2011), the children learn that a girl or boy leaves a cupcake in a brown paper bag inside the classroom. While the boy or girl is outside, another child helps the teacher tidy up and throws the paper bag in the trash. Responses to two FB questions were coded: the *morally relevant contents FB of the accidental transgressor* (what did they think was in the paper bag—trash [1 point] or a cupcake?) and the *morally relevant location FB* of the victim (where will the owner of the cupcake look for it—on the table [1 point] or in the trash)? Cohen's κ indicated high agreement between the two raters, $\kappa = 1.00$.

Mental Verbs of Certainty. In the mental verbs of certainty task (adapted from Moore et al., 1989), the children played a hiding game with stickers. The children were told that a sticker was hidden in one of two boxes (red and blue) and that they could ask two puppets (a lion and a bunny) to help them find the sticker. The puppets made claims regarding one of the boxes. During the practice trials, the lion would, for instance, say, "The sticker is in the red box," while the bunny would say, "It's not in the blue box." The child was asked to find the sticker. Practice trials were repeated until children answered correctly.

Across nine trials, the lion and the bunny used different mental verbs of certainty: know, think, and guess. For example, the lion would say that he knows the sticker is in the blue box and the bunny would say that he guesses that the sticker is in the red box. When the children picked the box suggested by the puppet using the stronger mental verb of certainty, 1 point was given on comparison trials (i.e., know vs. think; know vs. guess; guess vs. think). There was high agreement between two raters, Cohen's $\kappa = 1.00$. The comparison between guess and think proved to be difficult, and so we computed a competency score based on only the comparisons between know versus think and know versus guess.

Metacognition (Partial Ignorance). Children's metacognition of their own (partial) ignorance was assessed with a task by Kloo et al. (2017). The experimenter hides one of two toys in a box (children cannot see which one). She then asks whether the children know which toy is hidden (two trials). One point was given when children spontaneously stated that they do not know or admitted that they were guessing on a subsequent "know-guess" question. There was high agreement between two raters (30% double-coded), with Cohen's κ ranging from $\kappa = .89$ to $\kappa = .93$ for the two tasks.

Second-Order FB. Two tasks assessed children's understanding of second-order FB.

In the "birthday puppy" story (Waves 2 and 3; Sullivan et al., 1994), the children learn that Peter's mom wants to give Peter a small dog for his birthday. When Peter asks mom what he will get for his birthday, she tells him that he will not get a dog because she wants to surprise him. When Peter plays in the basement, he finds the dog. The children are asked two control questions: Does Peter know that he will get a dog for his birthday? And, does his mother know that he found the dog in the basement? The story then continues. Grandma calls and asks mom whether Peter knows what he will get for his birthday (Test Question 1) and what mom thinks Peter thinks that he will get (i.e., the second-order FB; Test Question 2). High agreement emerged between the two raters (30% double-coded), with Cohen's κ ranging from $\kappa = .92$ to perfect agreement for both test and control questions. One point was given when both test questions were correct.

The simplified second-order FB task (Wave 1; Coull et al., 2006) presents the story of Anna who is hiding Paula's teddy bear. What Anna does not know: While hiding the teddy bear under the blanket, Paula observes her. The children are asked three control questions: Does Paula know where the teddy bear is? Does Anna know that Paula knows? And where will Paula look for the teddy? The test question asks where Anna thinks Paula will look for the teddy (and why?). One point is given when control and test questions are correct. Cohen's κ for interrater agreement was 1.00 for all control and test questions.

Understanding Nonliteral Speech (Strange Stories). We tested children's understanding of intentional, nonliteral speech with a selection of the strange stories (Happé, 1994). These story problems tested children's understanding of a lie (a girl lies about the dog breaking a vase when it really was her), a figure of speech (metaphor; someone says that someone else has a frog in their throat), and a joke (calling a big dog an elephant). The children were asked why the story characters made the particular utterances and answers coded using a coding scheme inspired by White et al. (2009). Correct answers explicitly referred to the intention or desire motivating the utterance (e.g., "because she did not *want* to get punished by her mother"; lie story); partially correct answers were purely motivational (e.g., "so that she would not get punished"). For the Rasch analysis, responses were dichotomized and full credit (1 point) was given for correct and partially correct answers. Cohen's κ ranged from .80 to .84.

Control Measures

Verbal IQ (Wave 0, 48 Months). An estimate of children's verbal IQ was obtained at a prior measurement point (Wave 0; 48 months). The verbal comprehension index was computed based on two subtests of the German edition of the Wechsler Intelligence Scale for Children: the "general knowledge" subtest (word knowledge and retrieval) and the "similarities" subtest (verbal concepts and reasoning). Interrater agreement was $r(25) = .97$ for general knowledge and $r(25) = .96$ for similarities.

Inhibition (Wave 2, 5 Years). We used a Simon (Peter) Says task based on the work of Strommen (1973) to assess inhibitory control. The experimenter instructed the children to execute the actions that she performed if and only if she said "Peter says . . . ." There were two practice trials with corrective feedback and 20 test trials. The inhibitory (i.e., non-Peter) trials were used for analysis (score range: 0–10).

Missing Data

At Waves 1–3, 130, 124, and 120 children participated in the assessments; 114 participated during all three. Complete data for all ToM measures were obtained from 84, 92, and 106 children, respectively. In the Rasch scale analysis (conducted in Acer ConQuest), missing data are omitted per task. The Martin-Löf (ML) test was conducted using the R package ltm (Rizopoulos, 2006). This analysis requires case-wise complete data, and so only children with full data were included. Estimates of verbal IQ and inhibition were obtained from 121 and 107 children, respectively.

Results

Correct Solutions

Tables 1 and S-2 report the correct solutions for all ToM measures. Across waves, there was considerable variation: At Wave 1, correct solutions varied between 4.0% (second-order FB) and 85.9% (DD and KA); at Wave 2, performance varied between 8.6% (joke story) and 80.7% (contents FB; lie story); and at Wave 3, between 47.1% (second-order FB) and 94.3% (lie story). The lie story was the easiest task at Waves 2 and 3; the joke story was the most difficult one at Wave 2. At Waves 1 and 3, where the joke story was not included, the two versions of the second-order FB task were most difficult. Spearman's correlations between all items are reported in the online supplementary material (Table S-3).

Extending the ToM Scale

A unidimensional, one-parameter Rasch model was fitted to the data of each of the three waves, using ConQuest (Wu et al., 2007). For all three waves, all ToM measures fit the Rasch model: All but one of the *infit* and *outfit* mean square (MNSQ) item fit statistics were within the margin of 1 ± .20 (i.e., $0.80 \geqslant$ *infit* MNSQ $\leqslant 1.20$; the misfitting item was the metaphor story applied at Wave 3). Applying a stricter criterion than this most commonly used one (i.e., $0.85 \geqslant$ *infit* MNSQ $\leqslant 1.15$) resulted in finding two tasks with poor item fit: These were the mental verbs task at Wave 1 and the metaphor story at Wave 3. Given both tasks did not exceed the standard cut-off values for the infit, they were not excluded from further analyses. The item difficulties and reliabilities for the three scales are given in Table 1, and item difficulties and person abilities are also plotted in the three Wright Maps in Figure 1, which also displays the three test information functions. These show that information for high ability levels decreases across waves, suggesting that additional more difficult tasks need to be included in (advanced) ToM scales after the age of 6 years.

The 1-PL (Rasch) model did not fit the data from Wave 2 significantly worse than the 2-PL model, with $\chi^2(8)=15.37$, $p=.052$. Although the likelihood ratio test was significant for Waves 1 and 3, with $\chi^2(8)=25.31$, $p=.003$ and $\chi^2(8)=20.08$, $p=.017$, respectively, the Bayes information criterion was—across all waves—lower for the 1-PL than for the less parsimonious 2-PL model: 1,318.82 vs. 1,337.32 (Wave 1); 1,168.07 vs. 1,191.39 (Wave 2); and 1,147.82 vs. 1,170.90 (Wave 3). Additional support for the Rasch model comes from the ML likelihood ratio test (Glas & Verhelst, 1995). The ML test splits the item set into two (or more) subsets, and the null hypothesis states that the Rasch model is valid for all items jointly. Using two subsets (median split), the null hypothesis was not rejected across waves—$ML(24)=22.287$, $p=.562$; $ML(19)=15.03$, $p=.721$; and $ML(24)=26.12$, $p=.347$, for Waves 1–3, respectively.

Table 1 shows that the estimated difficulties for the six core tasks that were included across all three waves. With increasing age, the relative difficulties of these tasks decreased, and they discriminated at a lower ability level. Overall, our findings suggest that the extended ToM scale shows good properties, and that diverse ToM tasks (incl. measures of metacognition and of morally relevant FB) load on a single dimension. Figure 1 shows increasing task difficulties that are in line with the developmental sequence and indicates a good match between difficulties and person abilities, suggesting that the tasks included in the analysis are age-appropriate and developmentally sensitive.

Stability of Individual Differences Over Time

Individual differences were moderately stable, as suggested by the significant autocorrelations. The correlation between performance at Waves 1 and 2 was $r=.54$ ($p<.001$); it was $r=.48$ ($p<.001$) between Waves 2 and 3, and $r=.40$ ($p<.001$) between Waves 1 and 3. ToM scores were significantly associated with verbal IQ (*r*s between .37 and .45, all *p*s < .001) and inhibition (*r*s between .21 and .44, all *p*s < .01) (see Table S-4 in the online supplementary material for the full correlation table, as well as for the partial correlations).

A repeated-measures analysis of covariance revealed significant effects of time—both for the full scales and the subsets of core items (Tables S-5 and S-6). Planned contrasts revealed a significant difference in children's ToM performance between 5 and 6 years, $F(1, 97)=5.66$, $p=.019$, partial $\eta^2=.055$—but not between 4 and 5 years, $F(1, 97)=.04$. The same was true for children's performance on the core ToM tasks, $F(1, 97)=5.32$, $p=.023$, partial $\eta^2=.052$ (comparison, 5 vs. 6 years), and $F(1, 97)=0.13$ (comparison, 4 vs. 5 years). No significant interactions between time and children's general cognitive abilities emerged.

Performance on the second-order FB task—a classic measure of advanced ToM—was significantly predicted by earlier first-order ToM skills: A binomial logistic regression with the second-order FB task (Wave 3) as the dependent variable, and verbal IQ, inhibition, and first-order ToM performance (as measured by the core ToM tasks at Waves 2 and 3) as the independent variables, revealed a significant effect, $\chi^2(4)=11.75$, $p=.02$ (adjusted $R^2=.16$). Of all predictors, children's performance on the core ToM tasks at Wave 1 was, however, the only significant predictor (Table S-7).

Taken together, these findings show that individual differences in ToM are moderately stable, suggesting conceptual continuity between first-order and advanced ToM.

Table 1. Performance (Percent Correct), Difficulty, and Fit Statistics for First-Order and Advanced ToM Tasks Across Waves.

	Wave 1 (4 years)					Wave 2 (5 years)					Wave 3 (6 years)				
	n	%	Diff.	Outfit	Infit	*n*	%	Diff.	Outfit	Infit	*n*	%	Diff.	Outfit	Infit
First-order ToM															
Not own desire	128	85.9	−2.45	0.96	0.99										
Knowledge access	128	85.9	−2.44	0.91	0.97										
Not own belief	129	66.7	−1.22	0.97	0.99										
Contents FB	128	36.7	0.20	0.97	0.99	124	80.7	−1.69	0.97	0.97	121	85.1	−0.62	0.95	0.97
Explicit FB	124	43.6	−0.12	0.97	0.98	125	76.8	−1.43	0.87	0.92	121	86.8	−0.77	0.74	0.91
Real-apparent emotion	100	24.0	0.92	1.13	1.05	117	40.2	0.42	1.09	1.05	121	76.9	0.00	0.89	0.98
Mental verbs	128	36.7	0.20	1.22	1.17	126	44.4	0.21	1.10	1.08	117	52.1	1.31	1.16	1.09
Morally relevant FB (location)	121	37.2	0.18	0.95	0.96	108	66.7	−0.83	0.97	0.98	113	90.3	−0.60	0.97	0.91
Morally relevant FB (contents)	122	12.3	1.76	0.95	0.98	108	58.3	−0.43	0.92	0.93	113	85.0	−1.14	0.83	0.92
Metacognition											117	75.2	0.12	1.08	1.03
Advanced ToM															
Second-order FB (Coull et al., 2006)	119	4.2	3.00	1.12	1.00										
Second-order FB (Sullivan et al., 1994)						125	20.0	1.52	1.05	1.03	121	47.1	1.56	0.97	0.98
Strange stories															
Lie						114	80.7	−0.96	1.02	0.97	121	94.3	−1.06	0.64	0.97
Joke						70	8.6	3.19	1.09	1.05					
Metaphor											103	68.9	1.19	1.32	1.20
Reliability															
Maximum likelihood estimate person separation reliability			.52					.48					.32		
Weighted likelihood estimate person separation reliability			.46					.41					.27		
Expected a posteriori estimate based on plausible values reliability			.40					.41					.26		

Notes. ToM: theory of mind; FB: false belief. infit = *infit* MNSQ (= mean square) statistic.

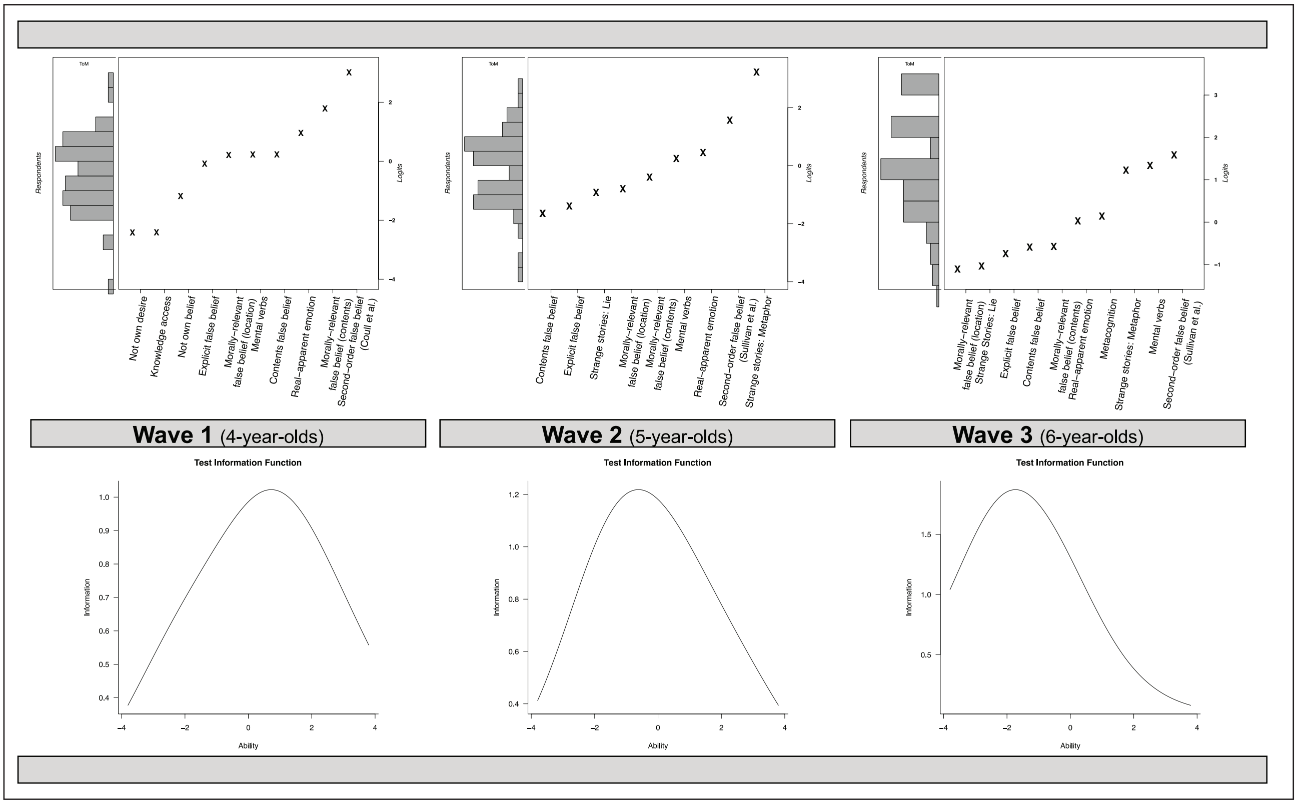

Figure 1. Wright Maps Across Waves (Top). The Difficulties of the Tasks Are Plotted on the Right; the Person Abilities on the Left. Test Information Function (Bottom). *N*s for Waves 1–3: 130, 124, and 120, Respectively.

Systematic Changes From Failure to Success

Systematic changes from failure to success with increasing age are considered one of the criteria that need to be met to conclude that a conceptual development is involved in a developmental progress (Wellman & Liu, 2004). To investigate whether there are subgroups of children who reveal systematic increases in performance over time, we used a *k*-means cluster analysis as a person-centered analysis of the six core (advanced) ToM tasks (i.e., the one assessed across all three waves). We started with a two-cluster solution and increased the number of clusters until no additional clusters with a case frequency >5 emerged. This analysis revealed eight clusters (Figure 2). All clusters revealed progressions over time that were either monotonic (Clusters C1, C3, or C6) or more pronounced between two of the three waves (Clusters C2, C4, C5, C7, and C8). Based on average cluster performance, we classified children as high performers (high performance across all waves; Cluster C1; $n=12$), low performers (low performance across all waves; Clusters C7 and C8; $n=17$), early bloomers (high performance at Wave 2, but not Wave 3; Clusters C4 and C5; $n=20$), or late bloomers (low performance at Waves 1 or 2, high performance at Wave 3; Clusters C2, C3, and C6; $n=68$).

A multinomial logistic regression analysis with the dependent variable "developmental pattern" (high vs. low performers; early vs. late bloomers) revealed significant effects of verbal IQ—$\chi^2(3)=15.04$, $p=.001$—and inhibition—$\chi^2(3)=9.52$, $p=.001$: Relative to the high performers, low performers revealed a significantly lower verbal IQ—odds ratio [OR] = 0.87, $t(1)=9.05$, $p=.003$—and a significantly lower inhibitory control—OR = 0.76, $t(1)=6.56$, $p=.010$. Whereas verbal IQ did not differ between high performers and early and late bloomers, both latter groups revealed a significantly lower inhibitory control—OR = .83, $t(1)=3.86$, $p=.049$ and OR = .81, $t(1)=5.22$, $p=.022$ for early and late bloomers, respectively. Taken together, these findings reveal systematic changes from failure to success in first-order and advanced ToM, which are related to verbal IQ and inhibition.

Measurement Invariance

Measurement invariance over time was investigated by assessing differential item functioning (DIF) for the six core ToM items and using the Mantel–Haenszel method, which is implemented in R package difR (Magis et al., 2010). Comparing Wave 1 to Waves 2 and 3, large DIF effects emerged for all items. Comparing Wave 2 to the other two remaining waves revealed significant and large DIF for contents and explicit FB, and morally relevant contents FB. Finally, comparing Wave 3 to the other two waves, significant DIF emerged for the real-apparent emotion and mental verbs tasks (see Table S-8 in the online supplementary material for the full results).

Discussion

The present study shows that the classic ToM scale (Wellman & Liu, 2004) can be extended by including a broad range

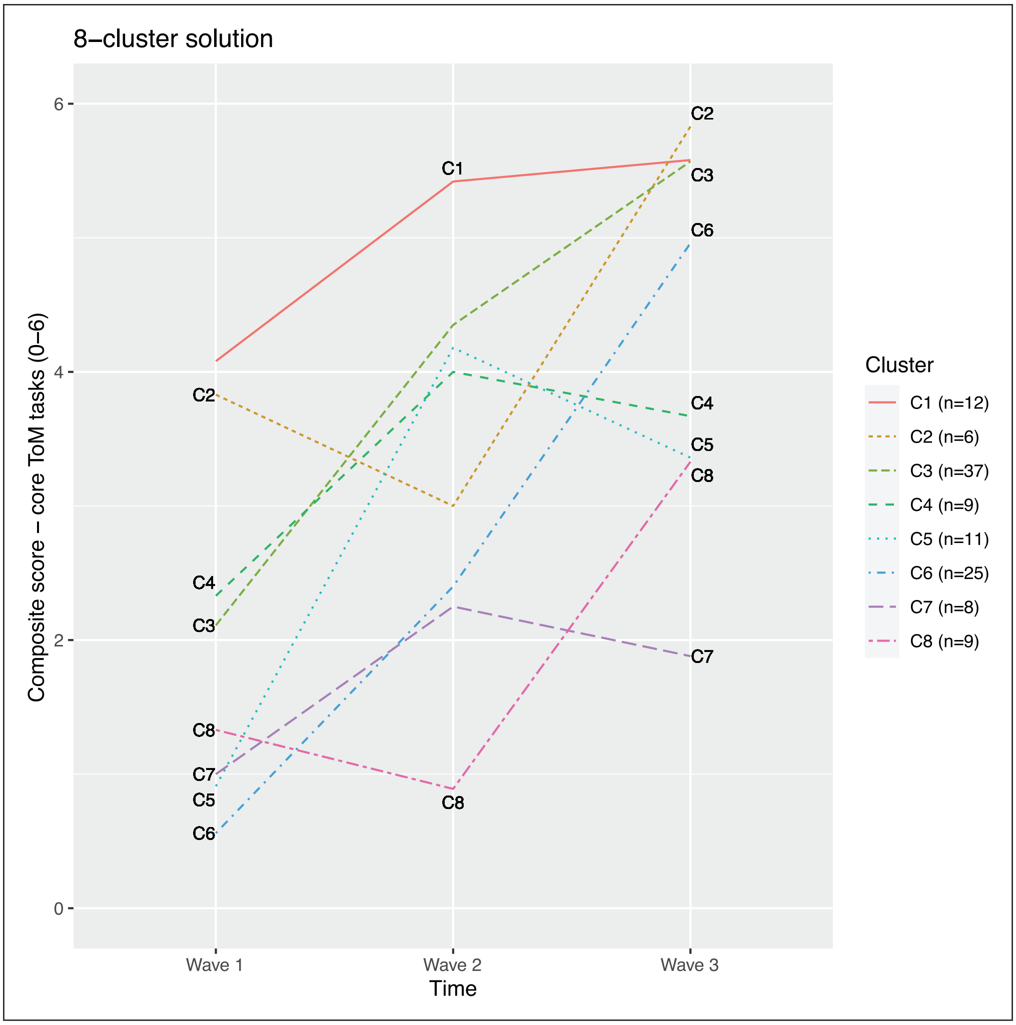

Figure 2. Results of a Cluster Analysis, Showing the Final Cluster Centers for an Eight-Cluster Solution. Children with Missing Data During One of the Waves Were Removed from the Analysis (*N* = 127).

of first-order and advanced ToM tasks. Our modeling of the developmental transition from first-order to advanced ToM suggests the presence of conceptual continuity between first-order and advanced ToM.

Our extended ToM scale included—in addition to (a selection of) tasks from the classic ToM scale—assessments of children's understanding of mental verbs of certainty, of morally relevant FB, and of second-order FB. These measures made up the core item set during all three waves. At Waves 2 (5 years) and 3 (6 years), we additionally included two strange stories that assessed children's understanding of nonliteral speech (lie and joke at Wave 2; lie and metaphor at Wave 3). The scale analysis revealed that the Rasch model fitted the data well, and individual differences were moderately stable over time, supporting the hypothesis that there is conceptual coherence in first-order and advanced ToM.

There was significant development in children's ToM, with performance differences being particularly pronounced between 5 and 6 years. This finding is in line with the literature, showing that many of the ToM concepts included in our extended scale (e.g., second-order FB understanding; metacognition regarding one's own ignorance) emerge during that period (Rohwer et al., 2012). The developmental progression between 5 and 6 years was, however, not independent of children's verbal IQ and their inhibition. This finding is in line with prior results that show that advanced ToM is rather loosely tied to children's age (Osterhaus et al., 2016), but closely associated with their executive functions (Devine et al., 2016).

Systematic differences between failure and success emerged between Waves 1 and 3: Our cluster analysis revealed that there was no group of children whose average ToM performance did not increase over the 2-year course of the study. Based on this person-centered analysis, we classified children as high versus low performers, or as early versus late bloomers (i.e., children with ToM development mainly between 4 and 5 years or 5 and 6 years). High verbal IQ and inhibitory control were associated with children being more likely classified as high performers, whereas low inhibitory control (but not low verbal IQ) predicted being classified as early or late bloomers. These findings are in line with theoretical accounts and empirical evidence regarding the involvement of language and inhibition in the emergence of ToM (Carlson & Moses, 2001; Milligan et al., 2007), and they suggest that low inhibitory control may mask an already existing mastery of ToM concepts (expression hypothesis, see e.g., Russell et al., 1991).

Several measures, including the metaphor and joke stories, and the lie story (at 6 years), did not discriminate well between children with and without advanced mindreading skills. Being a widely used measures of advanced ToM, this finding needs to be followed up and researchers should investigate the longitudinal associations between first-order ToM and the strange stories to establish that both require similar mindreading skills. Also, measurement invariance did not hold across all waves, with especially large DIF effects during Wave 1 (4 years). This is an important finding because it shows that researchers seeking to track developmental change over time must carefully select (and construct) appropriate tasks for which measurement invariance holds.

This study adds to the growing literature showing that children's development of mindreading skills is not confined to preschool years. Our extended ToM scale can be used to measure children's progress on broad ToM concepts, and to capture individual differences in a developmentally sensitive way.

Conclusion

Our extended ToM scale allows for the psychometrically sound and developmentally sequenced measurement of a broad range of first-order and advanced ToM concepts. The longitudinal stability of individual differences and the systematic changes from failure to success support the hypothesis that there is conceptual coherence not only between early implicit and later explicit ToM (Sodian et al., 2020), but also between preschool and more advanced ToM in middle childhood.

Acknowledgements

The authors would like to thank two anonymous reviewers for their helpful comments.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors are grateful to all participants for their friendly collaboration and support of this research. This research was supported by the German Research Council (DFG SO 213/27-1,2).

ORCID iD

Christopher Osterhaus https://orcid.org/0000-0002-1262-2427

Supplemental Material

Supplemental material for this article is available online.

References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesle.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032–1053. https://doi.org/10.1111/1467-8624.00333

Coull, G. J., Leekam, S. R., & Bennett, M. (2006). Simplifying second-order belief attribution: What facilitates children's performance on measures of conceptual understanding? *Social Development*, *15*(3), 548–563. https://doi.org/10.1111/j.1467-9507.2006.00340.x

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*, 758–771. https://doi.org/10.1037/dev0000105

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341–349. https://doi.org/10.1037/1040-3590.8.4.341

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). Springer.

Happé, F. G. (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

Hayward, E. O., & Homer, B. D. (2017). Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. *British Journal of Developmental Psychology*, *35*(3), 454–462. https://doi.org/10.1111/bjdp.12186

Hughes, C., & Devine, R. T. (2015). Individual differences in theory of mind from preschool to adolescence: Achievements and directions. *Child Development Perspectives*, *9*(3), 149–153. https://doi.org/10.1111/cdep.12124

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition*, *119*(2), 197–215. https://doi.org/10.1016/j.cognition.2011.01.006

Kloo, D., Rohwer, M., & Perner, J. (2017). Direct and indirect admission of ignorance by children. *Journal of Experimental Child Psychology*, *159*, 279–295. https://doi.org/10.1016/j.jecp.2017.02.014

Kristen, S., Thoermer, C., Hofer, T., Aschersleben, G., & Sodian, B. (2006). Skalierung von "Theory of Mind"-Aufgaben [Scaling of theory-of-mind tasks]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *38*(4), 186–195. https://doi.org/10.1026/0049-8637.38.4.186

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model—Some critical suggestions on traditional approaches. *International Journal of Testing*, *5*(4), 377–394. https://doi.org/10.1207/s15327574ijt0504_3

Lagattuta, K. H., Elrod, N. M., & Kramer, H. J. (2016). How do thoughts, emotions, and decisions align? A new way to examine theory of mind during middle childhood and beyond. *Journal of Experimental Child Psychology*, *149*, 116–133. https://doi.org/10.1016/j.jecp.2016.01.013

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

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

Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development*, *60*(1), 167–171. https://doi.org/10.2307/1131082

Osterhaus, C., & Koerber, S. (2021). Social cognition in and beyond kindergarten: The relation between first-order and advanced theory of mind. *European Journal of Developmental Psychology*, *18*, 573–592. https://doi.org/10.1080/17405629.2020.1820861

Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of advanced theory-of-mind tasks. *Child Development*, *87*(6), 1971–1991. https://doi.org/10.1111/cdev.12566

Peterson, C. C., & Wellman, H. M. (2019). Longitudinal theory of mind (ToM) development from preschool to adolescence with and without ToM delay. *Child Development*, *90*(6), 1917–1934. https://doi.org/10.1111/cdev.13064

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Development*, *83*(2), 469–485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25.

Rohwer, M., Kloo, D., & Perner, J. (2012). Escape from metaignorance: How children develop an understanding of their own lack of knowledge. *Child Development*, *83*(6), 1869–1883. https://doi.org/10.1111/j.1467-8624.2012.01830.x

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The "windows task" as a measure of strategic deception in preschoolers and autistic subjects. *British Journal of Developmental Psychology*, *9*(2), 331–349. https://doi.org/10.1111/j.2044-835X.1991.tb00881.x

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology*, *47*(5), 1239–1247. https://doi.org/10.1037/a0023899

Sodian, B., Kristen-Antonow, S., & Kloo, D. (2020). How does children's theory of mind become explicit? A review of longitudinal findings. *Child Development Perspectives*, *14*(3), 171–177. https://doi.org/10.1111/cdep.12381

Strommen, E. A. (1973). Verbal self-regulation in a children's game: Impulsive errors on "Simon Says." *Child Development*, *44*(4), 849–853. https://doi.org/10.2307/1127737

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, *30*(3), 395–402. https://doi.org/10.1037/0012-1649.30.3.395

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

Wellman, H. M. (2020). *Reading minds: How childhood teaches us to understand people*. Oxford University Press.

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

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*, *80*(4), 1097–1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. ACER Press.