The development of theory-of-mind and the theory-of-mind storybooks

Blijd-Hoogewys, Els Maria Arsene; Blijd-Hoogeweys, E.M.A.

CHAPTER 2

Measuring Theory of Mind in children Psychometric properties of the ToM Storybooks

Abstract: Although research on Theory-of-Mind (ToM) is often based on single task measurements, more comprehensive instruments result in a better understanding of ToM development. The ToM Storybooks is a new instrument measuring basic ToM-functioning and associated aspects. There are 34 tasks, tapping various emotions, beliefs, desires and mental-physical distinctions. Four studies on the validity and reliability of the test are presented, in typically developing children (n=324, 3-12 years) and children with PDD-NOS (n=30). The ToM Storybooks have good psychometric qualities. A component analysis reveals five components corresponding with the underlying theoretical constructs. The internal consistency, test-retest reliability, inter-rater reliability, construct validity and convergent validity are good. The ToM Storybooks can be used in research as well as in clinical settings.

This chapter is published as: Blijd-Hoogewys, E.M.A., Van Geert, P.L.C., Serra, M., & Minderaa, R.B. (2008). Measuring theory of mind in children. Psychometric properties of the ToM storybooks. *Journal of Autism and Developmental Disorders*, online version.

INTRODUCTION

From the beginning of the last century, research has been undertaken on the social empathy of children (e.g. Butterworth & Light, 1982; Piaget, 1929; Selman, 1980). However, this topic only attracted the full attention of developmental psychologists after Premack and Woodruff (1978) introduced the term Theory of Mind in their chimpanzee research. Under the flag of 'Theory-of-Mind' it has become one of the most prolific research areas in social developmental psychology. Theory-of-Mind (ToM) is the social cognitive ability to attribute mental states to oneself and others and to use these attributions in understanding, predicting and explaining behavior of others and oneself (Mitchell, 1997). ToM is also referred to as 'folk psychological abilities' or as 'mind reading skills'. It is a core human capacity needed to fully understand the social environment and for showing socially adequate behavior (Astington & Jenkins, 1995).

After the pioneering work of Premack and Woodruff (1978), research in normal ToM development proceeded with Wimmer & Perner (1983) who aimed their research at the understanding of wrong beliefs in young children. This was soon followed by studies in deviant ToM development. Concerning the latter, a great deal of research has been aimed at children with autism, starting with the studies of Baron-Cohen, Leslie and Frith (1985). They formulated the assumption that children with autism lack a ToM and that this deficit can explain a crucial part of the social impairment of these children. Since then a considerable amount of research has been undertaken in both typically developing children and children with autism (for a review see Wellman et al., 2001, and Baron-Cohen, 1989a, 2000, respectively).

The majority of ToM research in children focuses on the comprehension of false beliefs. A false belief (FB) is the ability of a child to predict the action of a second person, while the child knows that this second person has an incorrect belief about the situation. Well-known paradigms used to test this are the Maxi test, which is an unexpected transfer test (Wimmer & Perner, 1983), and the Smarties test, which is an unexpected content test (Perner et al., 1987). In the Smarties test, a child has to predict what a second person will say what is in the Smarties container, given that the child has seen that a pencil has been put in it (he holds a true belief) whereas the second person has not witnessed this (he holds an incorrect belief). Children only succeed on such tasks if they acknowledge that people act according to their own beliefs, even if those beliefs are, according to the child,

wrong: the second person will say that the container holds smarties (and not a pencil).

The mastering of FB is considered to provide stringent evidence of a mature ToM (Hala & Carpendale, 1997). As a result, the question of how and when children appreciate FB has moved centre stage in research on social cognitive development (Russel, 2005). In addition, FB comprehension appears to be a universal milestone that occurs around the age of four, across different cultures (e.g. Callaghan et al., 2005; Wellman et al., 2001). However, equating FB understanding with the possession of a ToM is too simplistic. ToM comprises far more than that, like for instance the understanding of desires and emotions (Astington, 2001). In addition, various ToM precursors are also involved. Already in the first and second year of life, a child develops socio-cognitive skills important for later ToM understanding, such as understanding intentional actions, engaging in pretend play, joint attention and imitation (e.g. Callaghan et al., 2005; Colonessi, 2005).

Lately, the focus of research has moved from specific FB understanding to a more developmental view (Wellman & Lagattuta, 2000; Steele et al., 2003) aiming at a wide range of ToM components that children develop between their second and sixth year (Wellman & Lagattuta, 2000). In this period, ToM evolves from a simple desire theory to a complete belief-desire theory, from true beliefs to false beliefs, and from the understanding of first-order beliefs to second-order beliefs. Which mechanisms underlie this development remains subject of discussion (for a review, see Astington & Gopnik, 1991; Hala & Carpendale, 1997; Leekam, 1993) (for a discussion, see Astington, 2001; Scholl & Leslie, 2001; Wellman & Cross, 2001; Wellman et al., 2001). Roughly, three viewpoints can be distinguished: the theory-theory view, the modular view and the simulation view. The theory-theory view assumes that the ability to form theories is an innate capacity, founded on a general learning mechanism. The child learns through hypothesis testing (Carruthers, 1996; Gopnik, 2003; Gopnik & Wellman, 1992; Perner, 1988, 1991, 1993, 1995; Wellman, 1990; Wellman & Bartch, 1988). The modular view assumes that ToM has a specific innate basis, part of which is modular and which is activated on the basis of maturation (Baron-Cohen & Ring, 1994; Fodor, 1983, 1992; German & Leslie, 2000; Leslie, 1987, 1992, 2000; Leslie et al., 2004). The simulation view emphasizes the aspect of putting oneself in another person's shoes, and thus of truly 'empathizing', which is the ability to recognize, perceive and feel directly the emotion of another person (Gallese,

2007; Gordon, 1992, 1996; Harris, 1992). Recently, a rapprochement seems to emerge between the different views on mindreading abilities, resulting in a more hybrid position combining both the theory-theory view and the simulation view (Keysers & Gazzola, 2007; Stueber, 2006).

Relatively regardless of the view one holds on the underlying nature of ToM, the majority of researchers broadly agree on a number of observable aspects or components that constitute ToM knowledge in children. In deciding which aspects to incorporate in the present study, we leaned heavily upon the work of Wellman (1990), not only focusing on core ToM components, like desires and beliefs, but also on associated aspects, like the recognition of emotions, perception knowledge and the difference between physical and mental entities. The result is a comprehensive test of ToM components and associated aspects.

Comprehensive ToM tests

Test psychologists recommend the use of comprehensive instruments composed of multiple tasks. Since aggregation favors broader applicability and reliability, such instruments can reduce standard errors and make measurements more reliable and valid. The total score of such a test is a compound score; that is, a score built of different parts. Research on ToM has shown that compound scores are more stable, because they average over multiple factors and lead to a more accurate measurement of the underlying skill (Hughes et al., 2000). In using such scores, a more adequate diagnostic procedure might be attained, which can help in studying the potential nature and causes of ToM differences in children (Hughes & Dunn, 1998). In addition to providing a single, quantitative measure of the level of ToM ability, it also allows investigators to compare different relevant ToM components or aspects in the same child and thus to discover how these aspects are related during the course of development.

In current research on ToM, such comprehensive tests are seldom used (for exceptions, e.g. Happé, 1994; Tager-Flusberg, 2003; Wellman & Liu, 2004). On the contrary, most research is based on single task measurements involving single aspects of ToM. These assessments may be quick and efficient, but provide no information about the nature and coherence of different aspects of ToM, and the stability of ToM ability over time. Examples of comprehensive ToM tests are the ToM battery of Happé (1994), the Tom-Test of Steerneman and colleagues (Steerneman et al., 2002; Muris et al., 1999), the ToM tasks of Tager-Flusberg (2003), and the ToM

tasks of Wellman and Liu (2004). The first three comprehensive tests incorporate both simple and more advanced aspects of ToM. The ToM battery of Happé (1994) incorporates first-order-belief tasks, first-order deception tasks; second-order-belief tasks and second-order deception tasks. The ToM tasks of Tager-Flusberg (2003) consist of three batteries tapping early (pretend & desire), middle (perception/knowledge, location-change FB, unexpected-contents FB & sticker hiding) and more advanced ToM aspects (second-order-belief, lies and jokes, traits and moral commitment). The Tom-Test (Steerneman et al., 2002; Muris et al., 1999) consists of three subscales tapping ToM precursors (e.g., recognition of emotions and pretense), first manifestations of a real ToM (e.g., first-order-belief and FB) and more advanced ToM aspects (e.g., second-order-belief and humor). The last comprehensive test, the ToM tasks of Wellman and Liu (2004), confines itself to simple ToM tasks only. The tasks tap various desires, diverse beliefs, knowledge access, content FB, explicit FB, belief emotion and real-apparent emotion.

The ToM Storybooks

Many ToM tests are aimed at testing school-aged children. However, ToM-problems often occur long before this age, as the CHAT (CHecklist for Autism in Toddlers; Baron-Cohen et al., 1992) and M-CHAT (Modified CHAT, Robins et al., 2001) illustrate by identifying potential ToM problems at the age of 18 months on. We did not have the intention to measure ToM functioning from this age on, but wanted to develop a comprehensive test that can be used to assess basic ToM functioning in an age range that is as wide as possible. The aspects we aim at are ToM aspects that normally develop in the preschool years, but that also show further refinements during the school age period. Therefore, in accordance with Wellman and Liu (2004), we decided not to include second-order-belief tasks or other more advanced ToM aspects. At the time of the instrument building, the test of Wellman and Liu had not yet been published; in contrast to the ToM battery of Happé and the ToM-test of Steerneman and colleagues. Since the latter two tests were considered too complex to be used in preschool children, we developed a new test, the ToM Storybooks. In the 2002 paper from the current authors (Serra et al., 2002) a preliminary version of the test was presented. The following requirements were set for the final version: the test must comprise a wide and representative range of ToM components, cover a broad age range in order to allow for direct comparisons between

children of different ages based on a continuous developmental trajectory and, finally, be optimally accessible and attractive to the youngest age range in particular, since that is the age range where the most rapid developments in ToM occur.

In this paper, we present four studies on the validity and reliability of the ToM Storybooks. The first study presents the construction of the new ToM Storybooks. The second study is aimed at the content validity of the test. The third study addresses the reliability of the test. Is the internal consistency of the test items sufficiently high? Are measurements repeatable, what is the test-retest reliability? What is the correspondence between raters evaluating the answers of children? The fourth study is aimed at the construct validity of the test. Is there convergent validity; does this test correlate highly with other tests that are known to measure ToM? As regards divergent validity: do the results obtained with this test differ sufficiently from tests not aimed at ToM, like an intelligence test and a language test? Since research has already shown that ToM results correlate positively with verbal intelligence scores (e.g. Hughes et al., 1999) and language scores (de Villiers, 2000; Happé, 1995; Tager-Flusberg, 2000), we expect the test to show a positive correlation with language and IQ-tests.

An important question regarding the validity of the ToM Storybooks is whether the test is able to distinguish typically developing children, from children with autism spectrum disorders. A related question is whether the results regarding validity and reliability obtained with typically developing children also hold for the children with autism spectrum disorders. The latter group is known to have ToM problems (Buitelaar et al., 1999; Dissanayake & Macintosh, 2003; Hill & Frith, 2004; Serra et al, , 2002; Yirmiya et al., 1998). We aimed to test children with PDD-NOS. If the test is able to distinguish the ToM functioning of children with PDD-NOS from that of typically developing children, it is by definition also suitable for distinguishing children with more severe impairments in ToM functioning, for instance children with a more severe pervasive disorder like an autistic disorder.

STUDY 1: DEVELOPMENT OF THE TOM STORYBOOKS

We wanted to develop a comprehensive ToM test that assesses a variety of ToM components and associated aspects, which develop during the

preschool years and also tend to further refine and increase during the early school years. The construction of the test is explained below.

Setting and participants

We tested 324 typically developing children that came from preschools, kindergartens and elementary schools. All children had a Dutch linguistic background, and did not have any language acquisition problems that could have hampered their performance on the ToM tasks (for the effect of language on ToM performance see for instance Garfield et al., 2001; Lohmann & Tomasello, 2003). Two Dutch language tests were used, depending on the age of the child. For 3-6 year olds, the Reynell was administered (test for receptive language comprehension; Van Eldik et al., 1997); and for 6-9 year olds, the TvK (Taaltest voor Kinderen, Language Test for Children; Van Bon, 1982) was used (subtests 'vocabulary' and 'sentence construction'). Language scores were available for 249 children (Reynell: n=170, TvK: n=79). Those children who did not receive a language test were older than 6 years and judged as having appropriate language skills by their teachers.

The sample consisted of 157 girls and 167 boys. The ages ranged from three up to and including eleven years (see Table 1 for the age distribution). Because the most rapid changes in ToM occur before the age of five years, there is an overrepresentation of young children[1].

Table 1: *Distribution in age groups of the typically developing children*

	Age (in years)						
	3	4	5	6	7	8-9	10-11	Total
Boys	32	31	31	31	15	14	13	167
Girls	29	24	32	26	16	12	18	157
All	61	55	63	57	31	26	31	324

[1] Also children older than five years were tested, in order to determine the upper-age limit of the test. In addition, testing older children makes comparisons between children with and without ToM problems easier.

Construction of the ToM Storybooks

Different components

Primarily based on Wellman's work (1990), core ToM components like desires and beliefs were included, but also emotions and associated aspects like the distinction between physical and mental entities, and understanding that seeing leads to knowing were included.

Emotion recognition is an important aspect, since discriminating and labeling facial expression of emotions lay the foundation for the ability to respond empathically to others (Feshbach, 1982). By the end of the first year, typically developing children respond differently to facial expressions of emotion in others (Baron-Cohen, 1994). At 20 months they use emotion words like happy, angry, sad, and scared (Flavell et al., 1993). At 3 years, they understand desire-based emotions (Yuill, 1984), and at 5 years belief-based emotions (Hadwin & Perner, 1991).

Beliefs and desires are considered core components of ToM. At 1,5 years children understand that other people have desires (Repacholi & Gopnik, 1997); at 2,5 years they have a desire theory (Wellman & Woolley, 1990); at 3 years a simple desire belief theory, they understand first-order beliefs (Bartsch, 1996; Wellman, 1990; Wimmer & Perner, 1983), and at 4 years a complete belief desire theory is established (Wellman, 1990). Four-year-olds have a representational understanding of beliefs (Gopnik, 1993; Gopnik & Astington, 1988; Perner, 1991). Finally, 4-year-olds can distinguish true and false beliefs (Hala & Carpendale, 1997; Wellman, 1990; Wimmer & Perner, 1983).

Concerning the associated aspects, during their second year children comprehend the difference between physical and mental entities (Wellman, 1990). At 3 years, they understand that seeing leads to knowing (Astington & Gopnik, 1991; Pillow, 1989; Pratt & Bryant, 1990).

The tasks used in the ToM Storybooks are based on tasks from former research. In table 2, an overview can be found of the origin of the tasks. The different components are ordered by the age children are able to successfully accomplish such tasks.

Task structure

We developed 34 tasks in total. Task examples can be found in Appendix A. The order of tasks and the number of questions per task are described in Appendix B.

Table 2: Origin of tasks used in the ToM Storybooks, ordered by age

Age	Sort of task	Task based on research from	Comparison with other comprehensive ToM tests
1	emotion recognition	Pons & Harris, 2000	Steerneman et al., 2002
2	desire resulting in action or emotion	Bartsch & Wellman, 1989; Wellman, 1990; Wellman & Bartsch, 1988; Wellman & Wooley, 1990	Tager-Flusberg, 2003; Wellman & Liu, 2004
3	mental physical distinction (including close impostors)	Wellman, 1990; Wellman & Estes, 1986	Steerneman et al., 2002
3	perception knowledge	Baron-Cohen & Goodhart, 1994; Pratt & Bryant, 1990	Tager-Flusberg, 2003
3-4	belief resulting in action or emotion	Bartsch & Wellman, 1989; Wellman, 1990; Wellman & Bartsch, 1988; Wellman & Wooley, 1990	Steerneman et al., 2002; Wellman & Liu, 2004
4-5	first order false belief	Perner et al., 1987; Steerneman et al., 2002; Tager-Flusberg, 2003; Wimmer & Perner, 1983	Happé, 1994; Steerneman et al., 2002; Tager-Flusberg, 2003; Wellman & Liu, 2004

Storybook structure

The 34 tasks follow each other in a natural way; they are interwoven in stories. The stories feature a main protagonist, named Sam. A coherent drawing style was used (for instance, Sam always wears the same cloths). Each task is illustrated with a full color picture. The drawings are enlivened by the use of toy doors that can be opened, magnetic emotion faces that can be placed on the characters, and patches of soft fur that can be caressed, if wanted[2]. Transitions between tasks are also accompanied by drawings and text, to keep the story going and to avoid too much switching between tasks.

There are six storybooks in total: How is Sam feeling?, Sam goes to the park, Sam goes swimming, Sam visits his grandparents, Sam at the farm, and Sam's birthday. The order in which the six books are presented to the child is partly fixed and partly variable. The administration starts with the book 'How is Sam feeling?' and finishes with the book 'It's Sam's birthday'. The order of the other four books is chosen by the child. By offering this choice, we intend to involve the child more in the testing, increasing the child's commitment and motivation. The four books can be considered parallel tests: they have an identical underlying structure and correlations between the different books are high (see Table 3).

Although we conceive of the storybooks featuring the character Sam as the default version of our ToM test, three additional versions of the test were developed, based on different protagonists. They are designed to be used in a time-serial design, preventing trivial learning effects that might result from mere repetition. In the present article, we will confine ourselves to the default version of the test, featuring Sam.

Table 3: *Correlations between the four parallel books within the ToM storybooks*

	Book 3	Book 4	Book 5
Book 2	.67	.79	.74
Book 3		.72	.74
Book 4			.77

[2] Non-graphical elements are distributed sparsely across the text; manipulation of these elements is not a necessary condition for answering the test. None of the children from clinical populations to which the test has been administered so far has shown any sign of disturbance or aversion for the non-graphical elements.

Testing procedure

The test takes 40 to 50 minutes, including a short break. The child sits at the left side of the administrator, so it can see the drawings clearly (the drawings are on the left side of the book, while the accompanying text for the experimenter can be found on the right side). The drawings remain in front of the child during the questioning in order to prevent mistakes due to memory requirements (in agreement with Charman et al., 2001).

Scoring procedure

Scoring items

The 34 tasks consist of 95 questions, namely 77 'test questions' and 18 'justification questions' in total. The test questions (for instance, Where will Sam look for his rollerblades? In the toy trunk or in the box?), can be considered a quick and less thorough method of testing, since they do not require justifications from a child. The answers to these questions are coded as correct or incorrect (1 or 0 points; maximum score=77). Because justifications are considered to better reflect the ToM knowledge of a child, most tasks also include such questions. Justification questions (for instance, Why will Sam look in the box?) result in 2, 1 or 0 points, depending on the correctness of the mental state terms spontaneously used by the child (maximum score=36) (for the scoring procedure see the right four columns of Appendix B).

In order to enable the standardized evaluation of the justifications, a category system has been developed, based on the category system used by Rieffe (1998), on different categories from Wellman (1990), and on an exploration of the empirical data (the elaborate category system can be requested from author EB). Two rules of thumb are followed in scoring the justifications. First, a justification can only be scored if the preceding test question is answered correctly. Second, the correctness of categories varies over the different types of questions. For instance, a desire answer can only be considered a correct category if it was used within a desire task and not within a FB task. Therefore, for each justification question, correct answer categories are determined. They are chosen from 21 formulated justification categories; in Appendix C definitions of these categories can be found.

ToM sumscore as an estimation of ToM ability

To assess the properties of the test items in estimating the ToM ability, a one-parameter logistic model (OPLM; Verhelst et al., 1995) was used. The key idea in OPLM, a unidimensional Item Response Model, is that for each item the probability of responding correctly to the item can be described by a particular monotonic increasing function of ToM ability. In OPLM, the particular functions of the items may differ in the item location (some ToM items are more difficult to master than others), and in the item discrimination (some items discriminate children better in their ToM ability than others). For the justification questions, with 3 response categories (2 points, 1 point or 0 point), a polytomous OPLM was used.

The OPLM showed a good fit for the 95 ToM items (77 test questions +18 justification questions), except for the three items of the inferred belief control task. For those items, a higher ToM ability did not result in a higher probability of giving a correct answer. Therefore, those items were eliminated from the ToM test. The OPLM of the 92 remaining items revealed a good fit for all items. All items contribute significantly to estimating the ToM ability. The correlation between OPLM ToM ability estimate and the ToM sumscore was 0.99. Thus, the ToM sumscore and the OPLM ToM ability estimate yield approximately the same results for ordering the children on their ToM ability. Therefore, we confine ourselves to a ToM sumscore; weighted values are not required. The testing with the ToM Storybooks results in a maximum total score of 110 points (ToM-total score), consisting of a maximum of 74 points for answers to the 'test questions' (3 inferred belief control questions are excluded) and a maximum of 36 points (=18*2) for answers to the 'justification questions'.

ToM quotient score

In addition to a total score, a ToM Quotient score (ToMQ) can be calculated. Norms for the ToM sumscores were obtained by applying a non-linear smoothing method over the raw data. Smoothness of the estimated curve is induced by weighing neighbouring observed scores (see for instance Simonoff, 1996; Härdle, 1991). A Fourier-series tenth-order polynomial based on a Loess smoothing technique (locally weighted least squares estimate) has been applied, which enables us to calculate the conversion curve. This curve enables us to determine the value of the smoothed curve at any possible age between 3 and 11 years. The conversion curve was calculated with the help of the TableCurve 2D programme (Systat, 2000).

Raw ToM sumscores were converted to Z-scores and converted to quotient scores (Wechsler, 1981). This is a standardized normed score, with an average of 100 and a standard deviation of 15 (for more details on the norming procedure of the ToM Storybooks, we refer to Blijd-Hoogewys et al., submitted b).

Conclusion

The ToM Storybooks have been developed with the aim of providing practitioners and researchers with a comprehensive ToM test assessing different basic ToM components and associated aspects. The test was administered to typically developing children.

The test consists of 34 tasks divided over six storybooks. It holds 74 test questions and 18 justification questions, resulting in a maximum total score of 110. The ToM sumscore can be considered a good estimation of ToM ability, as the OPLM results illustrated. Weighted values are not required. Also, a ToM quotient score can be calculated.

As far as typically developing children are concerned, the test focuses on the age range between three and six, given the rapid developments of ToM that occur in this period. However, the test has standardized norms and is applicable up to the age of 12. As a result, an 11-year-old child with ToM problems can be compared to a typically developing 5-year-old but also to a typically developing 11-year-old. Thus, the test is particularly suited for studies requiring age comparisons, based on the same instrument (e.g. cross-sectional research, assessment of clinical populations at various ages).

Criteria for subgrouping

Since ToM evolves over time, one expects the ToM total scores to increase with age. There is indeed a significant positive correlation between ToM total score and chronological age in the NT group (N=324, r=.76, p<.001) (see Table 4 for the ToM total scores over age).

The dependence of the scores on age poses a number of problems for the analyses of reliability and validity. Hence, where needed, analyses were carried taking into account an age correction or by using distinct age groups. In the group of 324 typically developing children, we distinguished three age groups. The subdivisions were made based on theoretical expectations

(expected levels of ToM functioning: low level, intermediate level and master level) and pragmatic grounds (approximate equal groups). The youngest (n=87; 3-4.5 years old) represents the age at which ToM development is at its beginning, at least as measured with the ToM Storybooks. The eldest (n=118; 6.5-11 years old) represents the age at which the ToM aspects measured with the ToM Storybooks is expected to have consolidated. The fastest growth of ToM is expected to occur in the intermediate age group (n=119; 4.5-6 years old).

Because this article discusses different psychometric studies, each with different sub-studies, we enclose an overview of the statistical analyses performed and their results (see Table 9). These results are discussed in more detail below.

Table 4: *ToM total scores over age*

Age	Number	Average ToM score		
		Test	Justification	Total
3	27	36.30 (8.06)	0.11 (0.42)	36.41 (8.07)
3.5	34	43.09 (10.30)	0.50 (1.11)	43.59 (10.97)
4	26	51.69 (9.55)	5.42 (6.18)	57.12 (14.32)
4.5	29	55.28 (7.38)	6.55 (4.59)	61.83 (10.89)
5	30	57.70 (7.28)	7.33 (4.16)	65.03 (10.46)
5.5	33	82.88 (6.29)	10.88 (4.75)	73.76 (9.95)
6	27	62.04 (7.49)	11.07 (5.52)	73.11 (12.21)
6.5	30	60.80 (6.94)	10.17 (3.97)	70.97 (9.96)
7	31	67.39 (4.74)	15.71 (4.23)	83.10 (7.84)
8+9	26	69.56 (4.36)	16.80 (4.49)	86.36 (7.96)
10+11	31	70.39 (4.53)	18.17 (4.28)	88.56 (7.71)

Note. Average ToM scores and corresponding standard deviations are reported.

STUDY 2: CONTENT VALIDITY

Our test, the ToM Storybooks, consists of different tasks on ToM components and associated aspects taken from literature. It contains tasks aimed at assessing five subtypes of abilities, namely emotion recognition, understanding of desires and beliefs, making the distinction between physical and mental entities, and seeing leads to knowing (compare Appendix A). To assess whether those subtypes are indeed present, a component analysis was performed.

We expected the subtypes to be correlated, and the correlation to depend on age. That is, the degree of differentiation is expected to be the largest at ages where ToM has rapid growth. Less differentiation, and hence greater correlations between subtypes, is to be expected in early stages of ToM.

Subjects and method

The analyses were based on the 324 typically developing children from Study 1. We calculated composite variables for the ToM Storybooks: for the different tasks, means were calculated over theoretically similar items. This resulted in 21 composite variables (between brackets are the number of test questions + number of justification questions) (for example of the tasks see Appendix A): emotion recognition (5+0) and emotion naming (5+2) (parts from the emotion recognition tasks); desire action (3+1), desire emotion recognition (5+1), and desire emotion naming (5+0) (parts from the desire tasks); standard belief emotion recognition (2+0) , standard belief emotion naming (2+1), standard belief action (3+1), changed belief action (1+1), not own belief action (1+1), not belief action (2+1), inferred belief (control) action (4+0), and (explicit) FB action (5+2) (parts from the belief tasks); mental physical senses (8+2), mental physical others (4+1), mental physical future (4+1), real imaginary (7+0), close impostor senses (4+0), close impostor others (2+1), and close impostor future (2+1) (parts from tasks aimed at the distinction between physical and mental entities); and finally, the variable seeing-is-knowing (1+1).

Statistical analysis

The scores of the three age groups (group 1: n=87, 3-4.5 years old; group 2: n=119, 4.5-6 years old; and group 3: n=118; 6.5-11 years old) were analyzed using a Simultaneous Component Analysis with Equal Pattern (SCA-P; Kiers & Ten Berge, 1994). SCA-P, which is a variant of Principal Component Analysis, estimates one pattern matrix for all three groups. As a result, the interpretation of the components (or factors) is equal for all groups, but the correlations between components and standard deviations of the component scores can differ across groups. To determine the number of components, the scree-test (Cattell, 1966), the eigenvalue-greater-than-one rule (Kaiser, 1960), and the substantive meaning of components, was used. Only minimum

loadings of .400 were considered. Finally, composite variables had to show adequate specificity for their components. Subsequently, internal consistency reliability was calculated for the components found.

Results

The scree plot of the SCA-P did not give a clear indication for 5 components. The eigenvalue-greater-than-one rule indicated that 7 components should be retained. We established the number of components on the basis of the substantive content of the components, determining whether increasing the number of factors still allowed the items of a factor to measure a clinical concept. A solution consisting of five components provided the best interpretation. This solution accounted for 53.8% of the variance. The pattern matrix was rotated using the oblique Promax rotation criterion. The resulting structure matrix revealed a reasonably simple structure of five components (see Table 5A): component 1 = belief action; component 2 = emotion recognition; component 3 = mental physical; component 4 = belief emotion; and component 5 = desire emotion. Two composite variables (from the original 21 formulated) also had loadings on other components, not being entirely specific, namely the composite variables 'mental physical senses' and 'close impostor future'. Two other composite variables did not fit this structure, namely desire action and seeing-is-knowing. The correlations between the components varied from .248 to .454 (see Table 5B).

Cronbach´s alphas, corrected for age, for these five components were calculated: component 1 (10 items, α=0.79), component 2 (4 items, α=0.47), component 3 (9 items, α=0.80), component 4 (25 items, α=0.62) and component 5 (14 items, α=0.61). Since the scores on the justification items depended on the child's answer on the related dichotomous items, justification items were not included in the calculation of the alphas, in order to avoid artifacts.

To assess the degree of differentiation in the three age groups, inter-factor correlations between the five components were computed within the three age groups. The correlations between the components were largest in the youngest group (average correlations, standard deviations of the correlations: M=0.47, sd=0.08), and comparable in the intermediate age group (M=0.26, sd=.013) and in the eldest group (M=0.26, sd=0.15).

Table 5: *SCA-P structure matrix and component correlation matrix*

5.A. Structure matrix with correlations between 5 components and 21 composite variables.

	Component				
	1	2	3	4	5
emotion recognition		.926			
emotion naming		.924			
standard belief emotion recognition				.833	
standard belief emotion naming				.831	
desire action			.418		
desire emotion recognition					.944
desire emotion naming					.956
mental physical senses	.459		.669	.494	
mental physical others			.570		
mental physical future			.455		
close impostor senses			.682		
close impostor others			.520		
close impostor future		.475	.523		
real imaginary				.460	
seeing-is-knowing					
standard belief action	.772				
changed belief action					
(explicit) false belief action			.430		
not own belief action	.807				
not belief action	.820				
inferred belief (control) action	.683				
(explicit) FB action	.560				

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaizer Normalization.
Note: correlations <.400 and >-.400 were omitted.

5.B. Component Correlation Matrix

Component	1	2	3	4
1				
2	.250			
3	.454	.377		
4	.388	.291	.388	
5	.248	.294	.322	.267

Conclusion

The component analysis resulted in a structure that largely corresponds with the underlying theoretical constructs from the test. The five components appeal to the five subtypes of abilities named in Appendix A), except for the composite variable 'seeing leads to knowing' which did not appear as a separate component. This is not surprising, since the composite variable consists of too few questions (only two). The internal consistency reliability is satisfying (although some Cronbach´s alphas are not >.70), since it concerns alphas on subparts of the test each containing a limited number of items, that are also corrected for age. The inter-factor correlations are consistent with the expectations: they are high in the youngest children implying that ToM abilities are not (yet) differentiated.

STUDY 3: RELIABILITY

In order to examine the reliability of the ToM Storybooks, we calculated the internal test consistency, test-retest reliability and inter-rater reliability. In addition, we examined the possibility of diminished test performance due to nuisance factors such as fatigue or boredom.

Subjects and method

For the internal test consistency, the data of the 324 typically developing children from Study 1 were used. For the test-retest reliability, a subgroup of 45 typically developing children (age 3-7) was tested again, with the second administration occurring two to three weeks later. We presume that ToM ability remains relatively constant when reassessed after such a short period. The test-retest reliability was also measured for children with PDD-NOS (n=18; age 5-9) (a subgroup from the clinical group that will be presented in Study 4, see Table 7), with the second administration after one week. In order to determine the inter-rater reliability of the justifications, the test results of a subsample of ten children were randomly chosen from both research groups (n=10 typically developing children and n=10 children with PDD-NOS). For the analysis of possible diminishing test performance at the end of the test, the data of the 324 typically developing children was used.

Statistical analyses

The internal consistency was established by means of a Cronbach's alpha. The test-retest reliability was established by means of a Pearson product-moment correlation coefficient. The inter-rater reliability was calculated on the basis of Cohen's kappa's. Five independent raters scored the justifications and the correlations between these five raters were calculated. This was done in two manners: a flexible manner by points awarded to the justifications (2, 1 or 0 points) (compare Appendix B) and a stringent manner by justification category chosen (compare Appendix C). To examine whether the test scores were affected by nuisance factors such as fatigue or boredom, it was checked if the results over the various storybooks showed a significant decline. Books 2 to 5 were considered, because they have a similar item structure (see Study 1). Since children could choose the order of the books, the average total score of the actual presentation of those books were compared. If nuisance played a part, the last presented book should result in a lower score than the first presented book.

Results

The internal consistency of the ToM Storybooks was good. After correction for the influence of age, Cronbach's alpha for the dichotomous items was .90. The test-retest reliability for the typically developing children was good (M_1 ToM-total score=59.91, $sd1$=18.46 versus M_2 ToM-total score=66.76, $sd2$=19.73; r=.86, p<.001). The children's scores rose significantly on the second administration (M=6.84, sd=10.33; paired samples t-test, p<.001). The test-retest reliability for the children with PDD-NOS was also good (M_1 ToM-total score=80.22, $sd1$=14.37 versus M_2 ToM-total score=79.67, $sd2$=15.67; r=.98, no significant difference). The inter-rater reliability was high (Cohen's Kappa=.97-.99 for the 0-2 points awarded, .81-.97 for the 21 categories). Concerning nuisance effects, no statistically significant decrease in total scores per book were found during the test administration; this applied for the total group as well as for the three separate age groups separately (for test performance from beginning to end of testing, see Table 6).

Table 6: *Test performance from beginning to end of testing*

	Book 2	Book 3	Book 4	Book 5
Age group 1 (3-4.5 years)	9.07 (3.37)	8.79 (3.38)	8.76 (3.45)	8.70 (2.85)
Age group 2 (4.5-6 years)	12.78 (3.44)	13.02 (2.57)	13.00 (2.97)	13.06 (3.55)
Age group 3 (6.5-11 years)	15.48 (3.25)	15.60 (2.41)	15.84 (2.89)	15.82 (3.01)
Total group	12.77 (4.19)	12.83 (3.84)	12.90 (4.15)	12.89 (4.24)

Note. Mean scores per book and standard deviations are depicted.

Conclusion

Based on the minimum standard for reliability of .70 (Nunnally & Bernstein, 1994, p. 265), the internal consistency (Cronbach's α=.90) of the total score was good (.90). This is an adequate value for a test aimed at young children and is consistent with findings from comparable research on standard and complex FB tasks (Hughes et al., 1999, 2000 obtained alphas of .83-.84; Muris et al., 1999 obtained alphas of .84-.92) and suggests that the different tasks measure the same underlying construct. Also, the test-retest reliability is good, both in typically developing children (r=.86) and in children with PDD-NOS (r=.98). This is consistent with findings from comparable research (r=.77: Hughes et al., 2000; r=.88: Muris et al., 1999). However, a significant increase in ToM total scores was found at the second measurement in typically developing children. Such a rise is not surprising, since it can be expected that young children learn from being tested (Grigorenko & Sternberg, 1998). The average score rise (M=6.86, SD=10.33) is of the same magnitude as those obtained with most standard psychometric measures on cognitive skills for young children. For instance, a difference of six IQ points can also be found in test-retest research with intelligence tests (e.g. Tellegen et al., 2003). A similar observation has also been reported in ToM research in typically developing children (Muris et al., 1999). The children with PDD-NOS did not show such a rise. They seemed not to have learned from their former experience. This finding may form an important point of attention in evaluating children with suspected ToM problems.

The inter-rater reliability of scoring the justifications is high (Cohen's Kappa>.80, namely .81-.97, even concerning the more stringent scoring criterion) (see also Charman et al., 2001; Muris et al., 1999). There were no differences in difficulty in judging the justifications of typically developing

children versus children with PDD-NOS. There was also no evidence for a statistically significant negative effect on the test scores due to increasing fatigue or boredom during the test administration.

STUDY 4: CONSTRUCT VALIDITY

We tested both the convergent and divergent validity of the ToM Storybooks. Concerning convergent validity, correlations with three similar tests were calculated. Concerning divergent validity, correlations with language and intelligence tests were calculated. The latter can be considered moderator variables in performance on ToM tests, but should not be considered to be equal to ToM. Despite their diversity, we do expect to find a positive relationship between ToM scores and scores on a language test, since ToM questions make a relatively strong appeal to lexical and syntactic knowledge (see for instance Garfield et al., 2001; Lohmann & Tomasello, 2003). We also expect a positive relationship with verbal IQ (Hughes et al., 1999).

Subjects and method

Children were referred to an outpatient clinic for child and adolescent psychiatry. After an extensive psycho-diagnostic and psychiatric examination (which included parent interviews and play contacts with the child), the children were diagnosed as having PDD-NOS (Pervasive Developmental Disorder Not Otherwise Specified) according to DSM-IV criteria (APA, 1994).

The clinical group consisted of 30 children with PDD-NOS. Their ages ranged from four up to and including eight years (see Table 7). There were 24 boys and 6 girls, resulting in a sex ratio of 4 to 1, which is the average sex ratio found in children with autism (compare Yeargin-Allsopp et al., 2003).

In order to check the validity of the clinical diagnosis, two additional tests were administered: the Vineland Adaptive Behavior Scales (VABS) (Sparrow et al., 1984; Dutch version: Researchgroup Developmental Disorders, State University Leiden, 1995) and the Children's Social Behavior Questionnaire (CSBQ; Luteijn et al., 2000; Dutch version: VISK; Luteijn et al., 2002; Hartman et al, 2007). The VABS is an interview in

which parents are questioned about the actual social behavior and skills of their child. We used parts of the expanded form of the VABS. For each child the discrepancy between the Vineland age equivalent (in months) and the chronological age (in months) was computed (VA-CA). The results showed that these children had large and negative discrepancy scores in receptive language, playing skills, interpersonal relationships and coping skills (see also Serra et al., 2002) as can be expected in children with pervasive developmental disorders. Their problems with expressive language and daily living skills (community) were less profound (Paul et al., 2004) (compare Table 7). The parents also filled in the CBSQ. This is a questionnaire in which parents report autism-related behavior. It can be used to facilitate selection of PDD samples for research purposes (Hartman et al., 2006). The CSBQ scores of our group are comparable to those known for children with HFA and PDD-NOS (compare Table V in Hartman et al., 2006) (total score, *M*=48, *sd*=18; 'tuned', *M*=10, *sd*=5; ' social', *M*=13, *sd*=5; 'orientation', *M*=8, *sd*=3; 'understanding', *M*=5, *sd*=3; 'stereotyped behavior', *M*=6, *sd*=3; 'change, *M*=2, *sd*=2).

Next, all children participated in an extensive psychological examination which included the assessment of intelligence (Wechsler Intelligence Scale for Children-Revised: Wechsler, 1974; Dutch version, 1986) and the level of language comprehension. Concerning the latter, two Dutch language tests were used, depending on the age of the child. For 3-6 year olds, the Reynell was administered (test for receptive language comprehension; Van Eldik et al., 1997); and for 6-9 year olds, the TvK (Taaltest voor Kinderen, Language Test for Children; Van Bon, 1982) was used (subtests 'vocabulary' and 'sentence construction').

Concerning convergent validity, two additional questionnaires and one test were included. The CSBQ (Luteijn et al., 2000) measures, among other things, ToM related knowledge, namely in the subscale 'difficulties in understanding social information'. The VABS questionnaire (Vineland Adaptive Behavior Scales questionnaire; Frith et al., 1994; Dutch translation: Hoogewys et al., 1999) consists of 32 theoretically derived items aimed at discriminating between social behaviors for which mentalizing (ToM) is essential (Interactive Sociability Scale, abbreviated as IS scale) or not (Active Sociability Scale, abbreviated as AS scale, concerning social behaviors that can be acquired without mentalizing). Both CSBQ and VABS questionnaire were administered for the clinical group (*n*=30 PDD-NOS, 4-8 years).

Table 7: *Test results of the children with PDD-NOS*

		Age (in years)						
		3 *n=2*	4 *n*=3	5 *n=4*	6 *n*=11	7 *n*=5	8 *n*=5	Total *n*=30
ToM-TB		41.00 (2.83)	42.67 (8.02)	55.75 (7.93)	75.73 (18.23)	71.60 (8.08)	80.80 (7.29)	67.60 (18.23)
VIQ		97.50 (10.61)	95.00 (24.43)	94.25 (14.45)	80.73 (14.53)	94.60 (11.13)	94.40 (11.57)	92.97 (13.48)
PIQ		116.00 (46.67)	104.00 (18.52)	109.50 (21.92)	94.32 (19.30)	108.80 (17.04)	107.20 (13.42)	103.32 (19.92)
VABS	Receptive language	35.50 (13.44) *-16.21 (19.91)*	33.00 (13.00) *-27.42 (14.65)*	43.33 (10.69) *-30.70 (12.22)*	45.00 (8.74) *-36.88 (9.87)*	48.00 (1.00) *-43.42 (3.17)*	44.20 (8.17) *-55.89 (9.43)*	43.25 (9.21) *-38.11 (14.18)*
	Expressive language	47.50 (9.19) *-4.21 (15.67)*	43.33 (14.19) *-17.08 (11.57)*	73.33 (27.79) *-0.70 (24.49)*	66.00 (19.21) *-15.88 (17.87)*	67.60 (23.83) *-23.82 (21.15)*	68.40 (18.06) *(-30.69 (17.30)*	63.75 (20.38) *-17.61 (19.12)*
	Community	44.00 (7.07) *-7.71 (13.55)*	41.00 (10.82) *-19.42 (8.47)*	63.33 (47.35) *-10.70 (19.71)*	67.60 (20.61) *-14.28 (20.77)*	70.20 (10.03) *-21.22 (6.84)*	74.60 (20.50) *-24.49 (20.94)*	64.32 (19.57) *-17.04 (16.87)*
	Interpersonal relationships	32.50 (13.44) *-20.71 (4.83)*	35.33 (15.37) *-26.08 (10.36)*	61.33 (47.35) *-2.37 (38.51)*	50.00 (23.09) *-32.58 (25.87)*	64.80 (22.90) *-26.62 (21.33)*	53.40 (33.63) *-45.69 (35.40)*	51.64 (26.72) *-29.08 (27.15)*
	Play & leisure time	19.50 (0.71) *-21.21 (9.78)*	37.00 (14.11) *-23.08 (11.47)*	51.33 (27.06) *-36.03 (17.47)*	47.50 (15.30) *-33.98 (15.35)*	54.60 (11.33) *-36.82 (11.31)*	63.80 (26.53) *-36.29 (16.87)*	48.96 (19.96) *-32.86 (16.33)*
	Coping skills	38.50 (17.68) *-16.21 (28.40)*	39.33 (10.41) *-14.08 (7.39)*	42.00 (17.35) *-24.70 (28.63)*	53.90 (21.40) *-29.38 (21.90)*	62.20 (11.68) *-29.22 (10.10)*	60.80 (16.72) *-38.29 (17.27)*	52.68 (18.27) *-27.86 (19.17)*

Note. VABS: means and standard deviations; every first row depicts VABS interview age equivalent; every second row (in italic) depi cts VABS interview discrepancy score (for each child the discrepancy between the Vineland age equivalent in months and the chronological age in months was computed for the different subscales).

Also a second ToM instrument was administered, namely the Tom-Test, a Dutch test that questions a wide variety of ToM aspects (Steerneman et al., 2002, 1999). In contrast with the ToM Storybooks, it also includes second-order-belief tasks. From the 30 children with PDD-NOS, 23 received the Tom-Test (age 4-8).

There were also four groups of typically developing children involved in Study 4. The first group is a subsample of 30 control children drawn from the 324 typically developing children from Study 1. They were matched on age and gender with the PDD-NOS group. This control group was used to make comparisons with the clinical group. The second is a subsample of 249 typically developing children (drawn from the group of typically developing children in Study 1; 3-9 years). For these children, language scores were available (Reynell: n=170, TvK: n=79; 59 boys and 48 girls). This control group was used to explore the relationship of ToM scores and language scores in typically developing children. The third is a subsample of 107 typically developing children (drawn from the 324 typically developing children in Study 1; 3-7 years). For these children, intelligence scores were available. They received a nonverbal intelligence test. Depending on the age of the child this consisted of the SON-R 2½-7 years (Snijders-Oomen Non verbal intelligence scale: Tellegen et al., 1998) or the SON-R 5½-17 years (Snijders-Oomen Non verbal intelligence scale - Revised: Snijders et al., 1988)[3]. This control group was used to explore the relationship between ToM scores and IQ scores in typically developing children. The fourth group is a subsample of 106 typically developing children (drawn from the 324 typically developing children in Study 1, 3-8 years; 54 boys and 52 girls). For these children, VABS questionnaire scores were available. This control group was used to explore the relationship between ToM Scores and VABS questionnaire scores in typically developing children.

Statistical analyses

With regard to convergent validity, we calculated Pearson product-moment correlations between the ToM Total score and the CSBQ subscales, the

[3] For pragmatic reasons, children with PDD-NOS were tested with the WISC. The division between verbal IQ and performance IQ can be very informative in children with autism spectrum disorders. Since the NT group also consists of children younger than 6 years, the WISC could not been applied and a non-verbal intelligence test was preferred.

VABS questionnaire and the Tom-Test. Divergent validity was tested by comparing the ToM Quotient scores with language scores and IQ scores by calculating Pearson product-moment correlations.

Results

The ToM scores of children with PDD-NOS are significantly lower than those of the matched control children (ToM total score: *M*=67.60, *sd*=18.23 versus *M*=77.23, *sd*=15.24, *p*=.001, one-tailed; ToM-Q score: *M*=85.10, *sd*=21.28 versus *M*=101.09, *sd*=13.79, *p*<.001, one-tailed). They had significantly lower scores on the mental physical tasks, the belief-action tasks, the belief-emotion tasks and the desire-action tasks. No significant differences were found for the emotion-recognition tasks and the desire-emotion tasks (see Table 8).

The correlations of the ToM Storybooks with other tests can be found in Table 9. The correlations of the ToM total score with the CSBQ subscales in children with PDD-NOS were negative and significant (p=.01, one-tailed): subscale 1 'not optimally tuned to the social situation' (r=-.26), subscale 2 'reduced contact and social interest'(r=-.26), subscale 3 'orientation problems in time, place, or activity' (r=-.60), subscale 4 'difficulties in understanding social information' (r=-.47), subscale 5 'stereotyped behavior' (r=-.39), and subscale 6 'fear of and resistance to changes' (r=-.41). The lower children with PDD-NOS scored on the ToM Storybooks, the more problems they exhibited on the CSBQ-subscales.

The correlations of the ToM Storybooks with the VABS questionnaire subscores are significant for the IS scale (*r*=.19 and *r*=.35, for respectively typically developing children and children with PDD-NOS, p=.01, one-tailed) and show a trend for the AS scale (p=.06, one-tailed, for both typically developing children & children with PDD-NOS). Thus, a higher ToM score implies higher sociability.

The correlation of the ToM Storybooks with the Tom-Test is high (*M*=47.09, *sd*=9.74 versus *M*=87.39, *sd*=11.36, scores of respectively Tom-Test and ToM Storybooks, *r*=.79, *p*<.001, tested two-tailed). Children with PDD-NOS evidence ToM problems on both the ToM Storybooks and the Tom test.

The correlation with language comprehension in typically developing children varies for the different language tests from .43 to .47 (p≤.001, tested two tailed; a common variance of 18 to 22%) (see Table 9). Concerning IQ, in typically developing children only a performance IQ was obtained. The

correlation with ToM-Q was.47 (p=.001, tested two-tailed; a common variance of 22%); while in children with PDD-NOS, there was no significant correlation with performance IQ (r=.07). However, the correlation of their verbal IQ with ToM-Q was.41 (p<.05, tested one-tailed; a common variance of 17%).

Table 8: *ToM results of children with PDD-NOS*

	Control group	**PDD group**	**Analysis**
Total score:			
Total ToM-score	77.23 (15.24)	67.60 (18.23)	MC, p<.001
ToM-Q score	101.09 (13.79)	85.10 (21.28)	MC, p<.001
Subscores:			
Emotion Recognition (0-14):	9.82 (2.62)	9.53 (2.84)	ns
Mental Physical:			
Real-mental items (0-24)	16.56 (3.42)	14.43 (3.82)	MC, p<.001
Real-imaginary items (0-8)	7.28 (1.15)	6.30 (1.64)	MC, p<.001
Close impostors (0-12)	8.81 (2.07)	8.13 (2.78)	MC, p=.01
Desires:			
Predicting action (0-5)	3.43 (1.05)	2.83 (1.18)	MC, p<.001
Predicting emotion (0-12)	7.94 (3.08)	8.03 (2.62)	ns
Beliefs:			
Predicting action (0-26)	16.31 (6.32)	13.03 (7.01)	MC, p<.001
Predicting emotion (0-6)	4.06 (1.43)	3.40 (1.52)	MC, p<.001

Note. MC=Monte Carlo analyses.

Conclusion

The results show that children with PDD-NOS evidence ToM problems. Children with PDD-NOS have problems with beliefs, both in predicting behaviors and emotions. In addition, they have problems on emotion recognition, real-imaginary, real-mental, close impostor, and desire-action tasks. These findings largely agree with the findings from Serra and colleagues (2002). Despite differences in p-values, the findings from both studies coincide. The only contrary finding is that beliefs used to predict actions were significantly more difficult for children with PDD-NOS than for typically developing children, whereas Serra and colleagues found the

opposite. The finding from the present study, however, is more consistent with clinical expectations.

The construct validity of the ToM Storybooks is good, both for the convergent and the divergent validity. Concerning the convergent validity, substantial correlations with ToM-related tests were found. The correlations of the ToM total score with the CSBQ subscales were good. Average negative correlations were found with all subscales. The highest correlations were found for the subscale 'difficulties in understanding social information', which can be perceived of as related to ToM skills, and for the subscale 'orientation problems in time, place, or activity', which can be perceived of as related to executive functions. It is known that executive functions are somehow linked with ToM development (e.g. Carlson et al., 2002).

The results from the ToM Storybooks also correlated with the VABS supplementary items from Frith and colleagues (1994). We found significant correlations with the IS Scale (requiring ToM) and a trend for the AS Scale (not requiring ToM) with the ToM Storybooks, for both the typically developing group and the PDD-NOS group. Our results agree to a large extent with the results of Frith and colleagues (1994), except that the latter found significant differences for the AS only in the normal control group and for the IS only in the autistic group. Purely speculatively, the differences in results can be due to the restricted use of FB measurements (Smarties test and Three Boxes test instead of a comprehensive ToM test) in a more seriously affected group (children with an autistic disorder compared to children with PDD-NOS in our research).

With regard to the convergent validity, the correlation of the ToM Storybooks with the Tom-Test of Steerneman and colleagues (2002) is as expected. The ToM Storybooks also have adequate discriminant validity. It can distinguish children with a normal ToM development from children with ToM problems, such as children with PDD-NOS. For future research, examining the applicability and discriminatory power of the ToM Storybooks, it is recommended to include children with an autistic disorder and other clinical groups, like for instance children with ADHD.

Finally, the correlations between children's scores on the ToM Storybooks and language acquisition tests are high (>.40). Also, correlations with IQ scores were inspected. The verbal IQ results of the children with PDD-NOS were somewhat lower than the normal sample, which is often seen in subsamples of children with autism (compare Joseph et al., 2002; Kraijer, 2004; Siegel et al., 1996). As regards the correlations with IQ scores, our research showed a significant correlation with PIQ for the

typically developing group (compare Muris et al., 1997; Muris et al., 1999; Carlson et al., 2002), but not for the PDD-NOS group. The latter group showed significant correlations with VIQ, consistent with findings of other researchers (r(230)=.43 in typically developing children, $p<.001$ in Hughes et al., 1999; $r(52)=.61$ in typically developing children and children with PDD-NOS, $p<.001$ in Muris et al., 1999). Final conclusions on the differences between these two groups cannot be drawn since different IQ tests were used. Due to the age limitations of IQ tests, different tests were used for the children with PDD-NOS in comparison with typically developing children. Moreover the PDD-NOS group was much smaller. As concerns future research on the relationship between IQ and ToM, the present authors recommend the use of a comprehensive ToM instrument, as was done in the present study and in studies of Hughes and colleagues, and Muris and colleagues.

Correlations between ToM Storybooks and IQ scores were notably high. However, this is not surprising. One could say that, if we look at intelligence in a broad way, comprehensive ToM instruments measure a specific aspect of intelligence, namely a kind of social intelligence. After all, these tests look into the logical reasoning of people and correlations of one type of intelligence with another are highly common. In addition, intelligence contributes to acquiring ToM skills, making it possible for children to understand connections between causes and results. In that view, comparison with IQ should perhaps not be considered as a test for divergent validity.

GENERAL DISCUSSION

This article presented the construction and validation of the ToM Storybooks. It is a comprehensive ToM test, measuring different basic ToM components, but also associated aspects. In Study 1 the construction of this test was discussed. The test holds 34 tasks, spread over six storybooks. A ToM sumscore and a ToM quotient score can be calculated. In Study 2, analyses showed an agreement between the underlying theoretical constructs and the components found through component analysis. Study 3 looked into the reliability of the test. Internal consistency, test-retest reliability and inter-rater reliability were found good. Lastly, Study 4 assessed the construct validity of the ToM Storybooks. Convergent validity, based on two

questionnaires and an additional ToM test, was good. The ToM-score had high correlations with language tests and IQ tests, as was expected.

It can be concluded that the validity and reliability of the ToM Storybooks comply with the requirements of an instrument of this sort. The separate findings are consistent with findings of other researches, but also agree with the more general findings of Wellman and colleagues on FB tasks (2001), which show that researchers can vary the tasks over an extended set of possibilities without influencing the performance of children. There is no indication that the medium in which ToM tasks is presented, in this case pictured storybooks, has affected the results in ways that reduce the test's reliability.

Additional validity research

Chapter five of this dissertation provides additional validity of the ToM Storybooks, namely criterion validity. In that chapter four quadrants of ToM functioning are hypothesized; and two of those quadrants are examined. The ToM Storybooks could distinguish well between children belonging to quadrant A or D. In Chapter six, there is mention of (ongoing) validity and replication studies.

A critical remark

The reliance of this kind of task on language comprehension with this kind of population, may lead to potential complications. Children with weak language comprehension undoubtedly will have more problems with successfully completing the test. The literature shows that there are strong relationships between language and ToM (Astington & Baird, 2004; Astington & Jenkins, 1999; de Villiers, 2000; Tager-Flusberg, 2000). In addition, many children with autism have language problems. In people with autism, ToM results are correlated to verbal mental age (Frith et al., 1991; Prior et al., 1990) and verbal skills (Happé, 1995). However, early research has shown that language problems do not contribute to mental state impairment, because children with for instance semantic language impairment do not show such problems (Leslie & Frith, 1988; Perner et al., 1989). On the other hand, the influence of language on ToM development should not be underestimated (Ruffman et al., 2003; Sparrevohn & Howie, 1995), also in testing. Language is a medium through which children learn

about beliefs (Astington, 2001). Reading storybooks, for instance, form a rich source of mentalizing information for children (Dyer et al., 2000).

Potentialities of the ToM Storybooks

The test includes a wider range of ToM aspects commonly tested. It includes not only tasks on first-order beliefs and desires, but also tasks on associated aspects such as the distinction between mental and physical entities. It is a comprehensive test consisting of tasks with different developmental challenges. The primary advantage of this test over existing batteries is that it targets skills that develop in typically developing children prior to the age of five, and further refine and increase during the early school years. The test, however, is applicable beyond the age of five; it has norm scores up to the age of 12 years and thus allows for comparisons between children of widely varying age, which makes it particularly appropriate for comparison with clinical groups in which ToM development is delayed. As a consequence, this test may have potential for a range of applications to both fundamental and applied work. Moreover, since this study covers a wider age range than is normally included in ToM research, valid comparisons between older children with ToM problems and their age mates with normal ToM functioning can be made. We like to remark that the older age group included in this study is not intended for discrimination between typically developing children, but between older children with clinical diagnosis. Since older typically developing children have, as a group, a smaller range in ToM total scores, a lower ToM score on these simple ToM tasks is very informative. Because of the use of simple ToM tasks and a motivating storyline, the test might also be useful in the field of intellectual disability, where autism spectrum disorders and related ToM problems are common. However, for future research it is advisable to include more complicated ToM tasks, such as a second-order belief task (see for instance Hughes et al., 2000) and a 'faux pas' task (Baron-Cohen et al., 1999), so that older children with more subtle problems can also be detected.

The test-retest correlations of the typically developing children suggested a small learning effect. As stated before, this is consistent with findings from Muris and colleagues (1999). Grigorenko and Sternberg (1998) recommended that this effect – the learning potential of individual children – be included in normal diagnostics. In that case, the pretest-posttest difference can eventually be considered an estimation of learning abilities that are, at

least in part, ToM specific. The absence of a comparable learning effect in specific groups of children, like we have found in children with PDD-NOS, could provide interesting information about the nature of ToM abilities in such children. In this line, further research on ToM might profit from dynamic testing– as opposed to static testing – where the learning potential of a child is quantified on the basis of his or her understanding and use of feedback given during testing (Grigorenko & Sternberg, 1998). Dynamic indexes can represent a quality step-up compared with static indexes (Fabio, 2005).

To conclude, one of the methodological strengths of the current test is that it has extended the limitations common in the majority of the researches done in the field of ToM. Most research has been undertaken in young children only (mostly up to 6 years, with a major focus on 3 to 4 year olds), has used only a few tasks (FB tasks, mainly single tasks) and considered small research groups (exceptions in the latter can be found in Charman et al., 2002; Hughes et al., 1999). The present research, aimed at constructing a new ToM Storybooks, used a wide range of tasks (not only FB tasks) and consisted of a substantial number of children over a wide age range. The test not only allows for comparisons on the basis of raw scores but standardized norms and norm scores are also available (Blijd et al., submitted a). In our opinion, the ToM Storybooks provide a comprehensive, valid and reliable instrument for researchers and clinicians who wish to measure Theory-of-Mind in young typically developing children, as well as children with an autism spectrum disorder from a broader age range.

Table 9: *Summary of the statistical methods*

Study	Characteristic	Typically developing	PDD-NOS[1]	Age correction	Results
Content validity		*n*=324, 3-11 years old		3 age groups: *n*=87, 3-4.5 years old *n*=119, 4.5-6 years old *n*=118, 6.5-11 years old	Simultaneous Component Analysis: 5 correlated factors
Reliability	Internal consistency	*n*=324, 3-11 years old		(α – (correlation ToM&age)²)/ (1-(correlation ToM&age)²)	Cronbach's α=.90
	Test retest reliability	*n*=45, 3-7 years old		no age correction applied	r=.86 ***
			n=22, 5-10 years old	no age correction applied	r=.98, ns increase
	Inter-rater reliability	*n*=10	*n*=10	no age correction applied	Cohen's Kappa =.81-.97
	Nuisance	*n*=324, 3-11 years old		3 age groups:	
				n=87, 3-4.5 years old	ns decrease/increase
				n=119, 4.5-6.5 years old	ns decrease/increase
				n=118, 6.5-11 years old	ns decrease/increase
Construct validity	Convergent validity:				
	- ToM SB[2] & CSBQ[3]		*n*=30, 4-8 years old	no age correction applied	SC[5] 3, r=-.60††; SC 4, r=-.47††
	- ToM SB & VABS -Q[4]	*n*=106, 3-8 years old		partial correlation	IS[6], r=.19†; AS[7], r=.13#
			n=30, 4-8 years old	no age correction applied	IS, r=.35†; AS, r=.24#
	- ToM SB & ToM test		*n*=23, 4-8 years old	no age correction applied	r=.79***
	Divergent validity:				
	- ToM SB & diagnosis	*n*=30, 4-8 years old	*n*=30, 4-8 years old	ToM Quotient scores	*M*=85.10< *M*=101.09***
	- ToM SB & language	*n*=249, 3-9 years old		ToM Quotient scores	Reynel,r=.47***;TvK[8], r=.43***
	- ToM SB & IQ scores	*n*=107, 3-7 years old		ToM Quotient scores	PIQ[9], r=.47***
			n=30, 4-8 years old	ToM Quotient scores	VIQ[10], r=.41* and PIQ, r=ns

Note. [1]PDD-NOS = pervasive developmental disorder n ot otherwise specified; [2]ToM SB= Theory of Mind Storybooks; [3]CSBQ= Children's Social Behavior Questionnaire; [4]VABS -Q= Vineland Adaptive Behavior Scales –questionnaire; [5]SC = subscal; [6]IS= Interactive Sociability; [7]AS= Interactive Sociability; [8]TvK= Language test; [9]PIQ= Performance IQ; [10]VIQ= Verbal IQ; ***p=.001, two-tailed; *p<.05, two-tailed; ††p<.01, one-tailed; †p=.01, one-tailed; #p=.06, one tailed.