THE TOM STORYBOOKS AS A TOOL OF STUDYING CHILDREN'S THEORY OF MIND IN FINLAND

Jussi Vesterinen
Master's thesis
University of Jyväskylä
Department of Psychology
April 2008

THE TOM STORYBOOKS AS A TOOL OF STUDYING CHILDREN'S THEORY OF MIND IN FINLAND

Author: Vesterinen, Jussi Tuomas
Master's thesis of psychology
Supervisors: E.M.A. Blijd-Hoogewys and Timo Ahonen
University of Jyväskylä, Department of Psychology
April 2008
35 pages

ABSTRACT

Background: Although false belief tests are valuable for scientific research on Theory of Mind (ToM), clinical and applied use require more comprehensive tests, containing multiple tasks on different aspects of ToM. The ToM Storybooks is such a comprehensive ToM test focusing on basic ToM components in children from three to five years old. It is a newly developed Dutch test. In this article, the value of a Finnish version is examined. **Method**: Forty two Finnish children (2-7 years old) were tested with a Finnish version of the ToM Storybooks. Their ToM knowledge was compared to that of a Dutch norm group. A subgroup of nine children was retested after 80 days. **Results**: According to paired comparisons there are no significant differences between the performances of Dutch and Finnish children. Internal consistency of the ToM storybooks was adequate. Children got better scores on their second testing. Finnish ToM scores were positively correlated with verbal intelligence and age. SDQ-Fin scales of prosociality and peer problems were not linked to ToM. **Conclusions**: The Finnish version of the ToM Storybooks can be applied to use with Finnish children. It gives versatile and reliable information and is able to differentiate between children on the basis of their ToM skills. Clinical use and further studies with the Finnish version of the test are encouraged.

Keywords: Theory of Mind, storybooks, Dutch, cross-cultural validation, reliability.

CONTENT TABLE

PREFACE

As an exchange student in Groningen, The Netherlands, I had the opportunity to learn about children's understanding of mind. I did not have much previous knowledge but I found the topic interesting since it was closely related to children's social and emotional development. After talking with E.M.A. Blijd-Hoogewys and Prof. Dr. P.L.C. van Geert, I decided to do a literature study on Theory of Mind (later abbreviated as ToM). They had been developing a new test called The ToM Storybooks which helps in understanding children's ToM skills. Later I was asked to translate the test into Finnish. It was an honor and I accepted. I was presented with the possibility of continuing with this topic in Finland by doing my thesis on The ToM Storybooks. All this experience encouraged me to take that challenge. Blijd-Hoogewys is my Dutch director and Prof. Dr. T. Ahonen is my Finnish director. They both deserve my gratitude especially due to their patience. ToM seems to be rapidly gaining interest in Finland. Many professionals in the area of psychology and psychiatry would like to have a contemporary tool for measuring different aspects of ToM especially in children with developmental disorders. Hopefully The ToM Storybooks will become a well-known, distinguished and available choice in the near future.

INTRODUCTION

What is Theory of Mind?

Theory of Mind is part of the social cognitive setting in psychology. The term 'Theory of Mind' or shortly 'ToM' was first used by Premack & Woodruff (1978) who were studying whether chimpanzees have mind-reading skills or not. ToM gained publicity among other researchers who soon began to study human children in order to learn how this ability is acquired (for reviews see Wellman, Cross & Watson, 2001; Baron-Cohen, 1989a, 2000).

ToM can be defined as the ability to impute mental states to others (Premack & Woodruff, 1978). It has also been referred to as ’everyday psychology’ (Wellman et al., 2001). Having a ToM enables us to recognize emotions, understand beliefs and desires, and predict and explain others' actions (Buitelaar, van der Wees, Swaab-Barneweld & van der Gaag, 1999). This makes ToM an essential skill for competent functioning and communication in everyday social situations (Astington & Jenkins, 1995). It has also an important role in emphasizing, understanding deception and allowing for self-consciousness and self-reflection (Frith & Happé, 1999; Howlin, Baron-Cohen & Hadwin, 1999).

There are different phases in the development of ToM, though there remains controversy about what those phases are and when they take place in child's development. It has been suggested that an important conceptual change in children's understanding of persons is taking place between the ages of 2,5 and 5 years (Wellman et al., 2001). This change has been characterized as a shift (1) from a situation-based to a representation-based understanding of behavior, (2) from a connections to a representational understanding of mind, or (3) from a simple desire to a belief-desire naive psychology.

Theories on Theory of Mind

Children with autism have serious social interaction problems. There are several hypotheses concerning the nature of their specific social-cognitive problems, the most influential ones being the ToM hypothesis and the rival affective or emotion recognition hypothesis (in Buitelaar et al., 1999). Some consider them as complementing activities (in Steerneman, Jackson, Pelzer & Muris, 1996). The biggest difference between them is whether ToM is seen as theory-like or not and perhaps the biggest thing in common is seeing ToM as a coherent and mentalistic skill.

According to the emotion-recognition hypothesis, people with autism fail to understand emotional expressions because they do not have the biologically based and normally innate capacity for it (in Buitelaar et al., 1999). This makes them fail in creating interpersonal connections when they are young and troubles the development of functions that are needed for interpersonal feeling. However, many studies have failed to replicate the results that have led to these conclusions.

The ToM hypothesis got the most attention. There are roughly three movements within the ToM field: the theory-theory, the modular view and the not-theory. There are also differences within these theories. Supporters of the theory-theory account see ToM as a highly theory-like conceptual framework that is very much like any scientific theory building and develops essentially by hypothesis testing (in Hala & Carpendale, 1997). Wellman (1990) argues that three important features of adults' understanding of mind, that can be found in scientific theories, are also apparent in even 3-year-olds' understanding. They are coherence (and interconnectedness of concepts), ontological distinctions (between mental and physical phenomena) and a causal explanatory scheme (explaining human action in terms of mental states).

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. Those who favor the nativist account emphasize innate factors in development instead of seeing the child as a scientist (in Hala & Carpendale, 1997). Perner's vision (1993) is partly theoretical but definitely mostly modular. He agrees that children make use of a theory and that the process of ToM development involves a dramatic theory

shift at around 4 years of age but he puts more emphasis to general cognitive ability to understand representations. He argues that young children conceive mental states and other representations simply as situations that correspond to a state of affairs (in Hala & Carpendale, 1997). Perner (1985) introduced the idea of dividing beliefs to first-order and second-order beliefs, which proved to be very helpful in measuring children's ToM and is used up till now. Another modular follower is Leslie. According to his early competence model, children understand persons' mental states because of a special Theory of Mind Mechanism (ToMM) that is activated early in development. There is also a Selection Processor (SP) that adjusts the functioning of ToMM by limiting it sometimes.

An example of a not-theory is the simulation theory. 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. Thus, this theory emphasizes the centrality of first-person consciousness (in Hala & Carpendale, 1997). Its advocates (like Gordon & Harris) deny the possibility that children's understanding of other minds would be theory-based and instead claim that this understanding takes place through a process of analogy. Several versions of simulation theory exist.

Children with Theory of Mind problems

As mentioned before, deficits in ToM abilities characterize individuals with autism. Autism is a complex disorder that affects many aspects of a child's functioning. Their social and communicative development is particularly disrupted, even in individuals who are of normal intelligence. They typically have rigid behavior patterns, obsessional interests and routines (APA, 1994; Howlin et al., 1999). Concerning ToM ability, they tend not to use mental-state terms in their spontaneous speech and they have difficulty distinguishing mental from physical entities (in Leekam, 1993). Across different studies only 20-50 % of the autistic children were found to pass first-order belief-understanding tests, compared to 65-80 % of non-autistic mentally retarded and 85-90 % of

typically developing control children (in Buitelaar et al., 1999). Studies on ToM in autism are important since they resulted in the most important information in ToM field, they make us better understand the problems associated with autism and also give us some idea of what life might be like without ToM (Baron-Cohen, 1989b; Buitelaar et al., 1999; Leekam, 1993). However, ToM problems are not unique to autism. They are also known in individuals with mental retardation, schizophrenia and in deaf children (in Yirmiya, Erel, Shaked, Solomonica-Levi, 1998). Also children with PDD-NOS (Pervasive Developmental Disorder – Not Otherwise Specified, a milder form of autism), ADHD (Attention Deficit Hyperactivity Disorder) and children with developmental language disorders (SLI, Semantic Language Impairment) are known to have ToM problems (in Buitelaar et al., 1999). What may be unique to autism is the severity of the ToM impairment rather than the impairment itself (Yirmiya et al., 1998).

The existence of ToM problems in other clinical groups, does not exclude the possibility that distinct elements (emphatic ability, various dimensions of cognitive ability, social relations, etc.) of this ability are differently impaired in various groups of individuals. For example, the studies on deaf children point to the importance of social learning or of an acquired element in ToM abilities, whereas studies regarding individuals with mental retardation point to the importance of cognitive faculties (Yirmiya et al., 1998).

Testing people's Theory of Mind skills

In doing research, one has to reflect on which tests and which research groups to involve. Concerning the tests to involve, various false-belief tests have frequently been used in measuring ToM abilities in autism. Perhaps the most common paradigm is the Maxi test that was introduced by Wimmer and Perner (1983). A variation of this paradigm is the Sally and Anne test (Baron-Cohen, Leslie & Frith, 1985). The test introduces Sally, who has a ball and puts the ball in a basket. She then leaves the room. Later on, Anne takes the ball out of the basket and puts it in a box nearby. So, the main character (Sally) does not know that the other character (Anne) has moved the object to a

different location. Then, the child is asked where the main character (Sally) will look for the object after coming back. To answer correctly, the test person (child) needs to comprehend that other people may have beliefs that are unlike the ones of the test person, that these beliefs may be false and that the character's actions are determined by his/her mental states. This is called a false belief test.

Another well known paradigm for testing false belief understanding is the Smarties (Perner, Leekam & Wimmer, 1987) or milk-carton task (in Yirmiya et al., 1998). There the child is presented with, for instance, a Smarties tube that contains something unexpected like a pencil. The participant is then asked to remember what the participant though to be inside the tube before knowing what it was, and what somebody else would say to be inside. Only 25 % of participants with high functioning autism seem to pass this test; so, it has good discriminating value (children who succeeded on this test had a minimal verbal mental age of 5,5 years old and a minimal chronological age of 11,5 years old; typically developing children succeed on this task around 4 years old) (in Baron-Cohen, Tager-Flusberg & Cohen, 1993). Besides that, these 25 % will also fail tasks that require second-order mental attributions. A task suitable for second-order ToM measuring could be the second-order belief attribution task developed by Wimmer and Perner (see Buitelaar et al., 1999). In other paradigms designed for studying ToM, the test person tries to understand various picture stories concerning mental states, mental physical distinctions, brain's function and deception (in Yirmiya et al., 1998).

To reiterate, there are many tests that can be used to measure children’s ToM, and false belief tests were the crown. There are three reasons that justify the use of such tests (in Wellman et al., 2001): First, with false-belief tasks it is easy and fast to assess if children understand that beliefs involve representations of reality and so can be mistaken. This is a very important feature of ToM understanding. Second, these tasks are very sensitive to early developmental changes which helped researchers to find out that even 4-year-olds have a surprisingly sophisticated ToM. Third, children with autism do very badly on these tasks, giving support to tasks' validity and highlighting the importance of the social insight that is needed in false belief tasks.

However, the ToM tests mentioned above also have limitations. A big problem about measuring children's ToM has been highlighted by psychometric testing theorists. They see a danger in measuring only single behaviors or focusing too much on single tasks (in Hughes et al., 2000). So instead of putting too much emphasis on false-belief tests alone, researchers should apply a task battery approach which means using a variety of tasks (Wellman et al., 2001). That way measurement errors average out and researchers get a broader picture of child's abilities that is also more reliable and valid. The Dutch ToM Storybooks is a comprehensive test that complies with those demands.

ToM has been measured with many different kind of tests and methods. Some differences between studies have created challenges to those who have compared the achievements of different studies. Children's performance in ToM tasks can be aided by some task manipulations. A child can be helped by increasing the child's participation by letting the child do the key transformations (e.g. hiding the ball in the Sally and Anne test), making the child realize the story's main character's mental state more obvious, reducing the influence of the contrasting real-world state of affairs and making the task involve explicit deception or trickery (Wellman et al., 2001). Furthermore, children from some countries perform significantly better than others.

Traditionally in research measuring development of ToM, test groups and control groups are not always balanced by verbal age but more often by chronological age. This can lead to underestimation of skills of some autistic participants who could pass the tasks presented to them if those tasks were not too verbally demanding. In that case the control group of the same age but normally developed verbal intelligence has the advantage over the test group. Those children with autism who do pass ToM tasks have been suggested to have higher verbal mental age, better skills in pragmatic language and better social functioning (in Leekam, 1993). However, this does not mean they do not have basic social and communicative difficulties.

The ToM Storybooks is a recently developed test for measuring many different aspects of children's ToM understanding. It is a more comprehensive test than most other tests and also more reliable because of the task battery approach. What makes it different from most tests designed for

that purpose is its diversity.

Objective of the study

This study introduces a newly constructed Dutch test, The ToM Storybooks (described in more detail later). This is the first small-scale pilot study of the Finnish version of the ToM Storybooks. The test results reported are those of Finnish and Dutch children of normal verbal intelligence. According to the null hypothesis there should be no significant differences between these two random samples. The effects of some potential background variables that could affect ToM scores are explored. In addition to nationality, the test results of Finnish children are expected to be connected to their verbal intelligence, age, peer problems and prosociality. Test-retest reliability and learning effects are analyzed by testing some children twice with moderate time difference between measurements. No significant improvement is expected. An important goal is the producing of information on attributes and usage of the Tom Storybooks as a potential tool for clinical practice in Finland.

METHOD

Sample

The sample of Finnish children was gathered from a kindergarten and a pre-primary school in Pyhäjärvi. Invitation to join the study was delivered to 52 children's families. In 42 families parents gave their consent and their child wanted to participate. There are 23 boys and 19 girls. Their native language is Finnish. The great majority of participants were 5 to 7 years old (Mean = 6 years 2 months, SD = 13 months).

Procedure

Children were tested one by one in a quiet place either in the kindergarten or the pre-primary school during those hours when children were available there. They were first presented with the ToM Storybooks and later on another day with a test on verbal intelligence (because of logistic reasons, only a subgroup was administered an intelligence test, namely 28 children). Nine children participated in the test-retest study: after an average of 80 days they were tested again with the ToM Storybooks. There was a half-time pause of five minutes during every test when children were encouraged to play or relax.

The administering of the ToM Storybooks took 26 to 45 minutes (not taking into account the half time pause). On the first testing round the average testing time was 33 minutes (N=42, SD=4). For the subgroup of nine children the first testing round was administered in 33 minutes (SD=4) and the second round in 29 minutes (SD=4). The second testing was done averagely 4 minutes faster (N=9, $p<0.01$, Wilcoxon Signed Ranks Test, 2-tailed).

Measures

A test on Theory of Mind – The ToM Storybooks

The ToM Storybooks is a Dutch psychological test made for getting information on the quality of a child's ToM skills and assessing whether these skills are developing with the child's age or not (Hoogewys, Loth, Serra & Van Geert 1998; for a preliminary version see Serra, Loth, van Geert, Hurkens & Minderaa, 2002). The test consists of six storybooks in which a main protagonist, named Sam, experiences all kinds of feelings, desires and thoughts. The child is asked a variety of questions about the protagonist's experiences. The questions are clustered in tasks. The tasks focus on ToM and associated aspects that children develop between the ages of three to six years old. They cover five components: 1) Recognition of emotion, 2) Distinction between physical and mental entities, 3) Understanding that seeing leads to knowing, 4) Prediction of behaviors and emotions from desires, and 5) Prediction of behaviors and emotions from beliefs (Blijd-Hoogewys, van Geert, Serra & Minderaa, under revision).

In each story the child is presented with an illustrated book that makes it easier to follow the stories read by the researcher. During the stories the researcher stops to ask the child some questions such as "Where will Sam look for grandpa?" and "Why is Sam looking under the table?" Giving the correct answer requires the child to take the perspective of the protagonist. Occasionally the child is also asked to connect the story character's mood to some pictures that represent different emotions like happy, angry, sad and normal.

For administering the test the researcher needs six storybooks, an empty score form and emotion cards. Based upon the six books, a total score is calculated. Subsequently a quantitative (max 76) and a quantitative + qualitative score (max 112) are possible. For the Dutch version, also a ToM quotient (abbreviated as ToM-Q) and a ToM age equivalent can be calculated (Blijd-Hoogewys, Van Geert, Timmerman, Serra & Minderaa, submitted a). ToM-Q is a normed quotient score with an average of 100 and a standard deviation of 15. Scoring the qualitative answers requires the researcher to be familiar with 21 different answer categories. In the current research, a Finnish translation of the ToM Storybooks version Sam was used (a revision of the test used in

Serra, Loth, van Geert, Hurkens & Minderaa, 2002). The author of this thesis, Jussi Vesterinen, has translated the test from the English version into Finnish. Children's answers were coded and sent to Blijd-Hoogewys who calculated the ToM total scores and ToM quotient scores with an Excel Visual Basic macro. The quotient scores are based on the normative data of Dutch children (N=324, 3-11 years old). They should not be used for calculating the scores of Finnish children but they can give some indication. The English score form was used for preventing children from seeing and understanding the answers in case they would try to have a look at the score form.

In order to chart ToM-strengths and ToM-weaknesses of a child, additional ToM sub-scores can be calculated, with the aid of the Dutch computer program. Questions that form these sub-scores are scattered in the six stories (see Appendix A for examples of tasks). Emotion recognition tasks (maximum score 14) require labeling the main character's current emotion and selecting it from the emotion cards. Emotions and actions are predicted on the basis of desires (17), beliefs (26) and false beliefs (9) that are either fulfilled or not fulfilled. Mental physical distinctions (24) require understanding if the situations are real and physical or mentally represented. In real imaginary distinctions (8) a child can be asked for example whether he can dream about dancing bananas or not. Close impostors (12) involve characteristics of physical objects that can be experienced in only few ways, such as smoke. Some tasks measure the understanding that seeing leads to knowing (3).

Other measures

Three subtests of the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R, Finnish version) were used for an index of verbal intelligence: ´Comprehension´, ´Vocabularity´ and ´Sentences´. A fourth subtest, ´Block Design´, was used as a sign of performance intelligence. Only 28 children were tested with WPPSI-R mostly due to the limited time.

The ´Strengths and Difficulties Questionnaire´ is a brief instrument designed for screening children who are at high risk for mental health problems (Goodman, Ford, Simmons, Gatward, & Meltzer, 2000). The ´Prosociality Scale´ and the ´Peer Problems Scale´ were chosen from a total of five scales. Parents were requested to answer to 10 claims concerning their children by choosing

one of three possibilities: 'not true', 'somewhat true' and 'certainly true'. The Finnish version of the test (SDQ-Fin) was used.

Statistical method

Independent Samples t-test, Pearson Correlation, Spearman Correlation and Wilcoxon Signed Ranks Test were used. Six Finnish children with low verbal IQ or ToM-Q were excluded from all comparisons so 36 children remained. The ToM Storybooks results of Finnish and Dutch children of normal verbal intelligence were compared in three ways: First, Finnish ToM-Q scores were compared to Dutch ToM-Q scores. Comparisons required a Dutch comparison group of the same age range as the Finnish children. From the 324 Dutch norm children, 259 children were of the same age range as the Finnish children. Their verbal or nonverbal IQ was at least 70 (not all children received an intelligence test, it was assumed that they had normal IQ's because they attended normal schools).

Second, Finnish children were paired with matching Dutch children and ToM total scores were compared. Third, the same operation was executed to compare ToM-Q scores. In total 23 Finnish children, who also had a verbal intelligence score (WPPSI-R), were chosen for paired comparisons. To make paired comparisons, 23 Dutch children were chosen, matched on age, gender and IQ-scores (where applicable). 12 of them were selected by age, gender and standardized verbal intelligence. 10 were selected by age, gender and nonverbal IQ (close to Finnish verbal IQ). One child was selected only by age. All 23 children chosen for paired comparisons belong to the previously selected group of 36 Finnish children. The Finnish sample had one girl more and one boy less compared to the Dutch sample.

The possible connection between ToM skills and verbal intelligence was explored by comparing the ToM total scores and verbal IQ's of those Finnish children whose verbal IQ was measured to be over 70.

RESULTS

The Finnish sample of 42 children was checked for outliers. The mean ToM-Q score was 87.02 (SD=22.86). This seemed to be low compared to Dutch mean ToM-Q score 99.9 (N=259, SD 18.26) (see Table 1). It was decided to exclude children with a low verbal intelligence (a WPPSI-R verbal IQ of 70 or less) or with a ToM-Q score under 50 from further analyses. As a result, 36 children remained in the group (see Table 2), with an average ToM-Q score of 92.94 (SD=16.12).

TABLE 1. Descriptive statistics on Dutch children

DUTCH CHILDREN	N	Mean	SE	SD
ToM total score	259	62.12	1.08	17.34
ToM-Q	259	99.9		18.26
Language comprehension IQ	170	108.62		12.60

Note: SE=Standard Error, SD=Standard Deviation

TABLE 2. Descriptive statistics on Finnish children

FINNISH CHILDREN	N	Mean	SE	SD	Min	Max
ToM total score	36	66.58	2.30	13.81		
ToM-Q	36	92.94		16.12		
Emotion recognition	36	7.58	0.39	2.31	2	10
Desires TOT	36	9.47	0.51	3.08	1	14
Beliefs TOT	36	12.94	0.67	3.99	4	19
False belief TOT	36	4.89	0.41	2.48	0	9
Mental Physical	36	14.64	0.58	3.47	6	18
Close Impostors	36	8.44	0.30	1.80	4	12
Real imaginary distinctions	36	6.75	0.22	1.30	4	8
Seeing leads to knowing	36	2.47	0.17	1.03	0	3
WPPSI-R verbal IQ	23	104.00		13.15		
SP Comprehension	23	11.30		2.03		
SP Vocabularity	23	11.26		3.14		
SP Sentences	23	9.30		1.77		
SP Block Design	23	10.04		2.57		
SDQ-Fin Prosociality	35	6.46	0.26	1.54	4	9
SDQ-Fin Peer problems	35	1.85	0.23	1.36	0	5

Note: SE=Standard Error, SD=Standard Deviation, Min=Minimum, Max=maximum.

Some Finnish children reached maximum sub-scores in false belief tasks, close impostors tasks, real imaginary distinctions tasks and seeing is knowing tasks. The most challenging tasks involved beliefs and mental physical distinctions.

TABLE 3. Pearson Correlations between ToM sub-scores in Finnish sample

	1.	*2.*	*3.*	*4.*	*5.*	*6.*	*7.*	*8.*
1.ToM total score	-							
2. Emotion recognition	.64**							
3. Desires TOT	.66**	.43**						
4. Beliefs TOT	.72**	.26	.34*					
5. False Belief TOT	.78**	.37*	.47**	.64**				
6. Mental Physical	.74**	.39*	.28	.39*	.51**			
7. Close Impostors	.72**	.59**	.39*	.44**	.38*	.54**		
8. Real Imaginary	.78**	.46**	.42*	.54**	.57**	.64**	.64**	
9. Seeing is knowing	.57**	.21	.34*	.38*	.39*	.43**	.55**	.43**

Note: ** $p < .01$ (2-tailed); * $p < .05$ (2-tailed)

Every sub-score was significantly correlated with ToM total score, and almost every sub-score with all other sub-scores (see Table 3). The internal consistency (Cronbach's alpha) was 0.85.

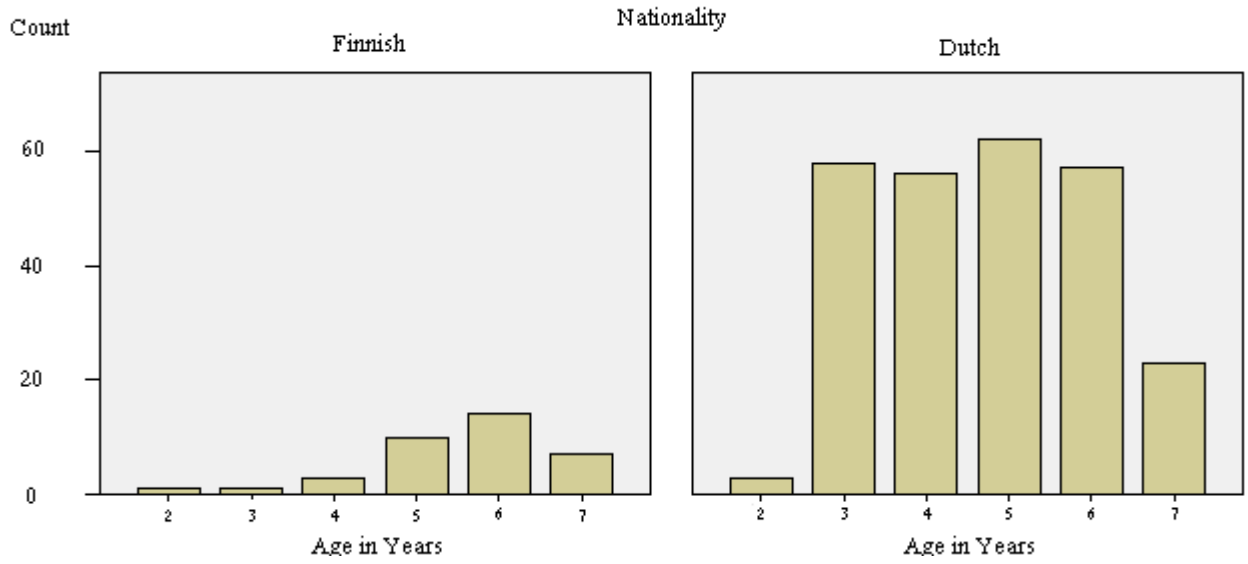

FIGURE 1. Ages of 36 Finnish and 259 Dutch children with ToM-score.

The average age of the Finnish group (N=36) was 73 months and the average age of the Dutch group (N=259) was 63 months (for the distribution, see Figure 1). The significant age difference

between samples (p<.001) points out that the effect of age must be controlled in ToM result comparisons. This can be done either by calculating ToM-Q scores or pairing parts of samples together.

The difference between ToM-Q scores of the Finnish (N=36) and the Dutch (N=259) samples was found to be significant (p=.039, 2-tailed, Independent samples t-test). Paired comparison between the ToM total scores of Finnish (N=23) and Dutch matching children (N=23) did not yield significant results (p=.46, Independent samples t-test). The Finnish average ToM total score was 69.7 and the Dutch average ToM total score was 72.1. The difference between ToM-Q scores was not significant (p=.37, Independent samples t-test). The average ToM-Q for Finnish sample (N=23) was 97.9 and for the Dutch sample (N=23) 101.6.

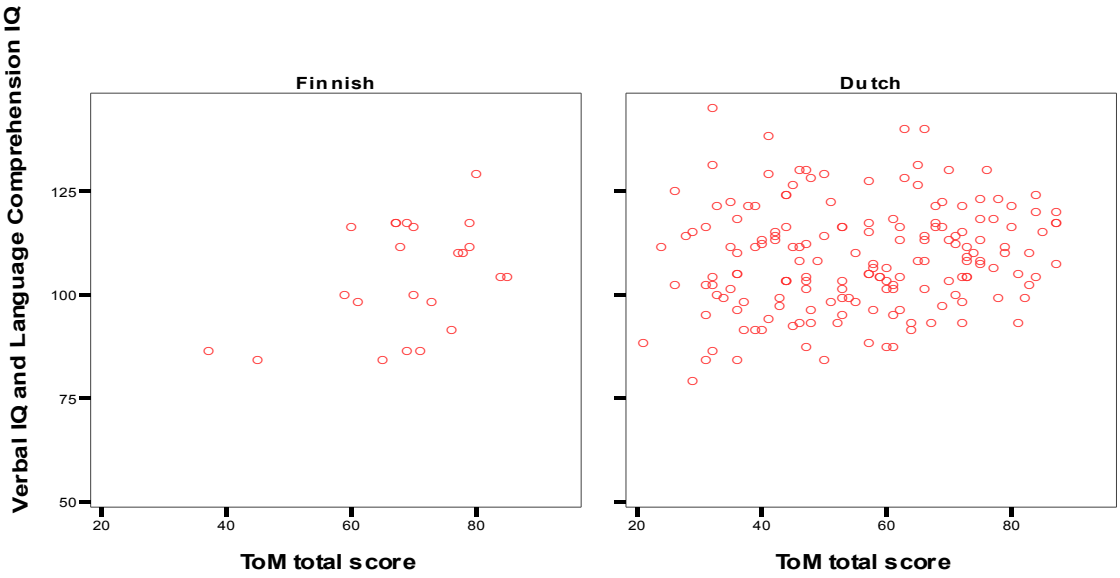

FIGURE 2. Scatterplot illustration on relationship of ToM and verbal intelligence.

There seems to be a more profound rising trend between ToM total score and verbal intelligence in the Finnish sample than in the Dutch sample (see Figure 2). Language comprehension of the Dutch sample had been measured using the Reynell (test for receptive language comprehension; Van Eldik, Schlichting, lutje Spelberg, van der Meulen & van der Meulen, 1997). It does not require children to use spoken language. This may have accounted for this difference in findings.

The relationship between the ToM total score and verbal intelligence was explored using

the Pearson Correlation. This correlation in the Finnish sample (N=23) was significant (r=.45, p=.03, 2-tailed). The Dutch sample (N=170) did not show significant correlation (r=.14, p=.07, 2-tailed) (See Blijd-Hoogewys et al., (under revision) for different results and more thorough analysis).

TABLE 4. Pearson Correlations between ToM sub-scores in Finnish sample

	1.	*2.*	*3.*	*4.*
1.ToM total score				
2. WPPSI - SP Comprehension	.33			
3. WPPSI - SP Vocabularity	.53*	.74**		
4. WPPSI - SP Sentences	.17	.37	.47*	
5. WPPSI - SP Block Design	-.13	.12	.18	.45*

Note: ** $p < .01$ (2-tailed); * $p < .05$ (2-tailed)

Possible correlations between ToM total score and WPPSI-R sub-scores (SP=standardized points) were explored (N=23) (see Table 4). ToM total score was significantly correlated only with vocabularity (r=.53, p=.01). Vocabularity correlated also with comprehension (r=.74, p<.001) and sentences (r=.47, p=.02). Additionally sentences correlated with block design (r=.45, p=.03) which is a nonverbal task.

Concerning the test-retest part, the second ToM testing occured approximately 80 days after the first one. Nine children took part in this study. The ToM total scores of both measurements were compared with a Wilcoxon Signed Ranks Test. The average ToM total score was 64.33 (SD = 16.27) for the first measurement and 75.78 (SD = 16.82) for the second measurement. These averages were significantly different (p=.015, 2-tailed). Spearman's correlation between testings was 0.38 (p=.32, 2-tailed). Eight children improved their scores and one child got a lower score on the second measurement.

Age was significantly correlated to ToM total score in both samples (p<.001). The correlation was 0.76 for the Finnish sample (N=36) and 0.75 for the Dutch sample (N=259). Gender was not significantly correlated with ToM total score in either sample (See Blijd-Hoogewys et al., (submitted a) for different results and more thorough analysis). ToM total score was not correlated to prosociality (r=.04, p=.81) or peer problems (r=-0.13, p=.47) (Pearson Correlation, 2-tailed).

DISCUSSION AND CONCLUSION

This study presented the ToM Storybooks and evaluated the Finnish version. The results gave us different answers on the question whether Finnish and Dutch children get through the ToM Storybooks similarly or not. Comparison between ToM-Q scores of all children suggested that the Dutch children were superior: they had better scores on this test. However, it should be noted that paired comparisons did not show significant differences between ToM total scores or ToM-Q scores. Therefore the null hypothesis is maintained: there are no significant differences between the two random samples. This is encouraging because Dutch norms were used and it can be assumed that the test favours Dutch culture.

ToM total score was positively correlated to verbal intelligence score, which is in concordance with findings from other ToM studies (e.g. Hughes, Deater-Deckard & Cutting, 1999). Age was strongly correlated with ToM total score, as expected. The sub-scores of the ToM Storybooks were correlated with ToM total score and also well within themselves. The internal consistency (Cronbach's alpha = 0,85) was adequate. This indicates that the different tasks measure the same underlying construct.

It seems that an excellent result in the test was not achievable for Finnish children. Some Finnish children were able to get the maximum points in four sub-scores out of eight. It should be noticed that these four sub-scores had fewer questions compared to the remaining ones. The ToM total score maximum is 112 points. The best Finnish score was 85 and the best Dutch score was 94, both existed of age and gender matched samples. The best Finnish ToM-Q was 125. On the basis of the lowest and highest sub-scores achieved by the Finnish sample, it can be concluded that the test is able to differentiate between subjects well.

The second measurement (after 80 days) of the average ToM score was significantly higher. Such a rise is not surprising, since it can be expected that young children learn from being tested (Grigorenko & Sternberg, 1998). The test seemed more familiar to children on the second testing round which was accomplished on average four minutes faster than the first round. Some

children seemed to remember some correct locations of hidden objects from the first time and this gave them some advantage which lead to a higher score. Still it is difficult to draw strong conclusions from a sample of nine children. Though, other ToM research has shown that such effects are common, also in ToM research (Muris et al., 1999).

Blijd-Hoogewys and her colleagues (under revision) have also found a significant rising (M=6.84 points, SD=10.33) on the children's scores when the second testing occurred two weeks after the first one (N=45, 3-7 years old, paired samples t-test, $p<.001$.) Interestingly children with PDD-NOS did not improve their score at the second measurement. 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.

ToM gives us tools for getting along with other people and understanding them. The SDQ-Fin scales were used to measure children's prosociality and possible peer problems. This study found no significant correlation between ToM total score and the SDQ-Fin scales, though the connection was theoretically considered highly likely. Perhaps the ten questions used from the SDQ-Fin were too imprecise for this purpose, parents' estimation skills were not accurate enough, or both tests do not measure the same underlying phenomena. Parents used both scales moderately: differences between minimum and maximum were five points but they could have been ten points. The average scores were 6.3 (SD = 1.5) for prosociality and 2.1 (SD = 1.5) for peer problems (N = 41). In a research of Obel et al. (2004) Finnish children's average scores were 6.6 (SD = 1.8) for prosocial behaviour and 2.4 (SD = 1.6) for peer problems (N = 727).

In the Dutch sample, the Reynell test was used to demonstrate the correlation between ToM total score and verbal abilities. This connection was weaker than the one found in the Finnish sample, using the WPPSI-R, even though the Dutch sample was considerably bigger. However, note that Blijd-Hoogewys and her colleagues (under revision) found different results in their more thorough analysis. They found correlations ranging from .43 to .47 between three different language comprehension tests and the ToM Storybooks (N=249, 3-9 years old; $p\leq.001$, 2-tailed; a common variance of 18 to 22%). Only a performance IQ was obtained in this sample.

In this study no significant connection was found between gender and ToM total score, which is not surprising taking into account the small number of subjects involved. Blijd-Hoogewys and her colleagues (submitted a) did find gender differences in their much bigger sample. First, they found that girls had slightly higher ToM total scores than boys (Independent samples t-test, p=.098) though the variances between sexes were considered equal (Levene's test, p=.749). So, on first inspection, it could be concluded that there were no gender differences. But, when different age groups were considered (*n*=87, <54 months; *n*=119, 54<78 months; *n*=118, ≥78 months), there were eminent significant differences between boys and girls. Gender differences were found in the oldest and youngest subgroup. Based on these results, separate norms for boys and girls were generated. Norms based on the total sample were also determined, since the overall difference between boys and girls was relatively small (about 0.15 of the standard deviation)

Limitations of the study

Children with low IQ's (lower than 71) were not included in this study, since ToM is correlated to intelligence: children with a mental retardation also have ToM problems. Six Finnish children were left out of comparisons based on their low performances on WPPSI-R and the ToM Storybooks. Maybe even more children would have been discarded if verbal IQ had been measured from all Finnish subjects (N=42). Thus, the sample of 36 qualified children is somewhat questionable even though the average ToM-Q was raised from 87 to 93. The sample of 23 was more controlled.

The ToM-Q results were lower in the Finnish sample. Perhaps Dutch norms should not be applied too seriously for foreign samples. Comparison of ToM total scores between the Finnish (N=36) and Dutch (N=259) samples was not sensible because average ages were too different. Also the distributions of ages in our sample created challenges for comparisons (see Figure 1). Including younger Finnish children would have been better.

The pairing between the Dutch and the Finnish children was based on best judgement of the author. The idea was to find as similar Dutch children as possible concerning verbal

intelligence, age and gender. In ten occasions nonverbal IQ had to be used instead of verbal IQ in Dutch sample. A different researcher might have come up with a different kind of pairing.

Probably the use of different tests affected different kind of trends between ToM total score and verbal intelligence (see Figure 2). Dutch average on the Reynell language comprehension test (N=170, M=108.62, SD=12.60) seemed higher than the Finnish average on the WPPSI-R verbal-IQ (N=23, M=104, SD=13.15). Off course both tests are not totally comparable, since the Reynell is no intelligence test. Also, maybe the language comprehension test was too easy. Children do not need to use spoken language in order to perform well on this test. It would be interesting to study if this is a matter of using spoken language or not. Maybe expressive language skills have stronger connection to ToM than just verbal intelligence and understanding of language.

Originally the order of tests presented to the children was planned to be randomised. Possibly presenting the WPPSI-R first might have activated vocabulary and reduced nervousness towards the ToM Storybooks and promoted better results in the latter. All testing situations were attempted to be interesting so that the children could concentrate about 45 minutes during both testing days. Some children expressed that they found the ToM Storybooks more interesting than the WPPSI-R. Many children wanted to be tested a third time.

My ideas and future directions

Some children got surprisingly low results on the ToM Storybooks. In the author's opinion certain characteristics seemed to have affected these performances: shyness, poor language skills, bad mood, low motivation, confusion and restlessness. On rare occasions some children made up funny or strange explanations to the test questions as if they did not take the test seriously or they had their own ideas about situations in the stories. These rare answers rarely got any points.

It has been supposed that children may find the correct answers to ToM questions through logical reasoning without much awareness of social factors (in Buitelaar et al. 1999). In that case involving justifications is important. For these kind of questions, explanations consisting of mental

states is an absolute requirement for success on the ToM Storybooks. Sometimes these explanations seemed to be scarce in the otherwise intelligent Finnish children. Situational explanations, not involving any mental states, were more commonly used than expected. Another common problem found in the Finnish sample was that children forgot the names of the emotions (sad, angry etc.). It is possible that the Finnish culture favours more situational language than language with mental state verbs compared to Dutch culture. Maybe some children gave meagre or short answers to the qualitative questions (justification questions) because they thought that the tester would understand without too detailed explanation.

It is not certain whether the test should be modified to be more suitable for Finnish use or not. For example, in one story the main protagonist, Sam, get's a jumper as a birthday present even though he did not want a jumper. Children are asked how Sam looks like (emotion). Children get points for answering ´sad´. In this study 23 children replied with ´sad´, 7 with ´normal´, 4 with ´surprised´, 1 with ´happy´ and 7 gave other answers. The typical way to act in this kind of situation in Finland may vary with age and other factors. Probably some adults would lie that they like the present to avoid hurting the feelings of the person who gave the jumper. Caring of another's feelings requires ToM. Probably this kind of things have been taken into consideration in designing the ToM Storybooks. The test was not made for adults but for children.

There are 21 categories for the qualitative answers (justifications) used in the Dutch version of the ToM Storybooks. The administer of the test uses a qualitative handbook for scoring the test according to these categories. The handbook has not been translated into Finnish. It is possible that different language might result in different categories. Exploring the need for changing categories and translating a Finnish version of the handbook are recommended. This would call for an extensive study involving considerably more children (like 100 children).

For the present, no children with special needs have been tested with the ToM Storybooks in Finland. Studying them would give precious information both on the Finnish version of the test and on the aspects of ToM in Finnish children with special needs. Especially people working with children with autism would benefit from detailed information on children's ToM skills.

REFERENCES

American Psychiatric Association (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: American Psychiatric Association.

APA (1994). *DSM IV Diagnostic and Statistical - Manual,* 4th Edition. American Psychiatric Association: Washington, D.C.

Astington, J.W., & Jenkins, J.M. (1995). Theory of mind development and social understanding. Cognition and Emotion, 9, 151-165.

Baron-Cohen, S. (1989a). Theory of mind and autism: A fifteen year review. In S. Baron-Cohen, H. Tager-Flusberg & D. Cohen (Eds.), Understanding other minds. Perspectives from developmental cognitive neuroscience. Second edition. New York: Oxford University Press.

Baron-Cohen, S. (1989b). The autistic child's theory of mind: a case of specific developmental delay. Journal of Child Psychiatry, 30 (2), 285-297.

Baron-Cohen, S. (2000). Theory of mind and autism: a fifteen year review. In S. Baron-Cohen, Tager-Flusberg & D.J. Cohen (Eds.). Understanding other minds: Perspectives from developmental cognitive neuroscience (pp. 3-20). Second edition. Oxford: Oxford University Press.

Baron-Cohen, S., Leslie, A.M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition, 21, 37-46.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D.J. (1993). The impairment of ToMM: some issues. In S. Baron-Cohen, H. Tager-Flusberg & D.J. Cohen, Understanding other minds. Perspectives form autism, pp. 102-105. Oxford University Press.

Blijd-Hoogewys, E.M.A., Van Geert, P.L.C., Serra, M., & Minderaa, R.B. (under revision). Measuring Theory of Mind in Children. Psychometric Properties of the ToM Storybooks.

Blijd-Hoogewys, E.M.A., Van Geert, P.L.C., Timmerman, M.E., Serra, M., & Minderaa, R.B. (submitted a). Norming the ToM Storybooks: A comparison between two methods.

Blijd-Hoogewys, E.M.A., & Van Geert, P.L.C. (submitted b). Discontinuous Paths in the Development of Theory-of-Mind. A nonlinear dynamic growth modeling approach.

Blijd-Hoogewys, E.M.A., Van Geert, P.L.C., Serra, M., & Minderaa, R.B. (in preparation). Temporal patterns in the development of theory of mind. A longitudinal study in children with PDD-NOS

Buitelaar, J.K., van der Wees, M., Swaab-Barneveld, H., & van der Gaag, R.J. (1999). Theory of mind and emotion-recognition functioning in autistic spectrum disorders and in psychiatric control and normal children. Development and Psychopathology, 11, 39-58.

Frith. U., & Happé, F. (1999). Theory of Mind and self-consciousness: What is it like to be autistic? Mind & Language, 14 91), 2-22.

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. British Journal of Psychiatry, 177, 534-539.

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124*, 75–111.

Hala, S., & Carpendale, J. (1997). All in the mind: Children's understanding of mental life. In S.Hala (Ed.) The development of social cognition (pp. 189-239). Hove: Psychology Press.

Hoogewys, E.M.A., Loth, F.L., Serra, M. & Van Geert, P.L.C., (1998). ToM Takenboek [ToM Story Books]. Groningen: University of Groningen, internal document.

Howlin, P., Baron-Cohen, S., & Hadwin, J. (1999). Teaching children with autism to mind-read. A practical guide. Chichester: Wiley.

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test-retest reliability for standard and advanced false-belief tasks across a wide range of abilities. Journal of Child Psychology and Psychiatry, 41, 483-490.

Hughes, C., Deater-Deckard, K., & Cutting, A. (1999). "Speak roughly to your little boy"?: Gender differences in the relations between parenting and preschoolers' understanding of mind. *Social Development (*Special issue on "Relationships and children's understanding of mind"), *8,* 143-160

Leekam, S. (1993). Children's understanding of mind. In M.Bennet (Ed.), The Child as Psychologist. An introduction of social cognition (pp. 26-61). Hemel Hempstead: Harvester Wheatsheaf.

Muris, P., Steerneman, P., Meesters, C., Merckelbach, H., Horselenberg, R., van den Hogen, T., & Van Dongen, L. (1999). The TOM Test: A new instrument for assessing theory of mind in normal children and children with pervasive developmental disorders. *Journal of Autism and Developmental Disorders,* 29, 67-80.

Obel et al. (2004) The strengths and difficulties questionnaire in the Nordic countries. European Child & Adolescents Psychiatry (suppl 2)

Perner, J. (1993). The theory of mind deficit in autism: Rethinking the metarepresentational theory. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), Understanding other minds: Perspectives from autism (pp. 112-137). Oxford: Oxford University Press.

Perner, J., Leekam, S., & Wimmer, H. (1987). Three-year-old's difficulty with false belief: The case for conceptual deficit. British Journal of Developmental Psychology, 5, 125-137.

Perner, J., & Wimmer. H (1985). 'John thinks that Mary thinks that…'. Attribution of second-order beliefs with 5-10 year old children. Journal of Experimental Child Psychology, 39, 437-471.

Premack, D., & Woodruff, G. (1978). 'Does the chimpanzee have a theory of mind?' Behavioural and Brain Sciences, 1, 515-26.

Serra, M., Loth, F.L., van Geert, P.L.C., Hurkens, E., & Minderaa, R.B. (2002). Theory of mind in children with 'lesser variants' of autism: A longitudinal study. Journal of Child Psychology and Psychiatry, 43, 1-16.

Steerneman, P., Jackson, S., Pelzer, H., & Muris, P. (1996). Children with social handicaps: An intervention programme using a theory of mind approach. Clinical Child Psychology and Psychiatry, Vol.1(2): 251-263.

Van Eldik, M.C.M., Schlichting, J.E.P.T., Lutje Spelberg, H.C., Van der Meulen, S.J., & Van der Meulen, B.F. (1997). Handleiding Reynell Test voor Taalbegrip (2e dr.) [Manual for the Reynell Language Apprehension Test]. Nijmegen: Berkhout/Lisse: Swets & Zeitlinger.

Wellman, H.M. (1990). The child's theory of mind. Cambridge, ma: mit Press.

Wellman, H.M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. Child Development, 72, 655-684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103-128.

Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation and normally developing individuals. Psychological Bulletin, 124 (3), 283-307.

APPENDIX

Appendix A: The Theory of Mind Storybooks: example tasks

Before beginning the test, the child is presented with drawings of five facial expressions (happy, scared, angry, sad, and surprised); there was also a neutral (just OK) face. The child was asked to provide labels with the faces in order to be sure that he/she recognized each emotional expression (see also Hadwin, Baron-Cohen, Howlin & Hill, 1996). If the child did not know or made a mistake, the experimenter gave the appropriate label. After practicing the emotions, the actual test begins.

There are 34 tasks (also see Appendix C); they can be divided in five groups.

1. Emotion recognition (maximum of 14 points)

There are five emotion recognition tasks: happy, scared, angry, sad and surprised. The child is presented with five situational descriptions. It has to choose the appropriate face and provide the correct emotion label. To avoid a response bias, the presentation order of the faces varied.

Example task (see figure 1): *'Sam has won shooting marbles. He has won the most beautiful marble.'* Questions: *1) Choose the face that matches. (emotion recognition), 2) How does he look? (emotion naming), 3) How come Sam is feeling happy?*

2. The difference between physical and mental entities

1) Mental-physical distinction (maximum of 24 points)

Pairs of real-mental contrasts are used in which the child has to compare two characters that have corresponding objective and subjective experiences. The child has to compare real situations with pretending, dreaming, thinking about things, and remembering things. The (justification) questions and item sequence were counterbalanced.

Example task (see figure 2): '*Sam, mummy and Sparky are going to the park. First, they are going to the pond. Sam gives bread to the ducks. And then mummy too. Sam's friend, John, can't go to the park today. John is sick and is lying in bed at home. John pretends to give bread to the ducks.'* Questions: *1) Who can really see the bread with his eyes? John or Sam? (mental physical senses), 2) How come... [Sam/John] can really see the bread with his eyes? 3) Who can really give the bread to the ducks now? John or Sam? 4) John plays. He pretends to feed the ducks. Can the mummy of John really give that bread to the ducks too? (mental physical others), 5) Who cannot save the bread now and give it to the ducks tomorrow? John or Sam*? (*mental physical future*).

2) Real-imaginary distinction (maximum of 8 points)

Questions are asked about real items and imaginary, non-existing items.

Example task: '*John and Sam are eating their sandwiches. 'John', says Sam, 'Listen. I know a fun game. I am going to ask you strange questions.'* Questions: *1) Do yellow bananas exist? 2) Do dancing bananas exist? 3) Can you think of yellow bananas? 4) Can you think of dancing bananas?*

3) Close impostors (maximum of 12 points)

Close impostors are physical objects that do not posses all characteristics of real objects. Real physical objects, like for instance chairs, have three characteristics, namely behavioral-sensory evidence, public existence and consistent existence. Close impostors can only be perceived in one modality and cannot be touched or acted upon. There are two tasks: one task is on smoke, the other is on a nasty smell.

Example task (see figure 3): '*Sparky, the dog, is rolling in the mud. 'Yak Sparky, you smell bad', says Sam. 'It stinks!*' Questions: *Can Sam touch the smell with his hands? Can Sam smell the smell? (close impostor senses) Can mummy smell it too? (close impostor others) How come mummy can smell it ... [too/not]? Can Sam save the smell in a box and smell it again tomorrow?(close impostor future)*

3. Perception knowledge (maximum of 3 points)

Only one task is involved. Questions are asked about the connection of seeing or not seeing something and knowing or consequently not knowing something (a subtest that was also included in the batteries of Tager-Flusberg, 2003).

Example task (see figure 4): *'Today, it is Sam's birthday. He is five. In the room there are two gifts on the table: a little parcel and a big box. Lisa, his sister, is allowed to look in the box, Sam however, can only touch the box'*. Questions: *1) Who knows what is in the box? Sam or Lisa? 2) Why does ... [Lisa/Sam] know what is in the box?*

4. Desires (maximum of 17 points)

The knowledge of desires allows one to predict both emotions and actions. Both sorts of tasks are incorporated into test items where desires are either fulfilled or not fulfilled.

There are five tasks on desire-emotions (wanting and getting/ not getting/ getting something else, and not wanting and not getting/ getting).

Example task: '*Come along Sam and Sparky', says mother, 'we are going home.' On the way home, Sam sees the ice cream man. He wants an ice cream. 'Mother, can I have an ice cream?', he asks. 'Off course', says mother and Sam gets a great ice cream.'* Questions: *1) Choose the face that matches. (desire emotion recognition), 2) How does he look?(desire emotion naming), 3) How come Sam is feeling... [emotion]?*

There are three desire-action tasks. Example task: *'They are at John's house. But John has hidden himself. Sam wants to go swimming and John has to come along to the swimming pool. He goes to look for Sam in the cellar. He opens the door. And yes! There is John.'* Questions: *1) What will Sam do now? 2) Why is he going ...[repeat previous answer]?*

5. Beliefs (maximum of 34 points)

Questions are asked about fulfilled or not fulfilled beliefs. These tasks, like desire tasks, can be used to predict both emotions and actions.

There are two belief-emotion tasks. Example task: *'Sam thinks his swimming trunks are on the chair. Sam goes to look on the chair. But there he finds a chicken!'* Question: *1) Choose the face that matches. (standard belief emotion recognition), 2) How does he look? (standard belief emotion naming), 3) How come Sam is feeling… [emotion]?*

There are eight belief-action tasks. They are all first order belief tasks: on standard belief, changed belief, inferred belief, inferred belief control, not belief, not own belief (or diverse-belief), explicit false belief and false belief (change-of-location, see figure below) tasks.

Example task (see figure 5): '*Grandpa and grandma are paying Sam a visit. Sam gets rollerblades from grandpa and grandma. He's very happy with the present. Sam puts the rollerblades in the toy trunk. Then, he goes upstairs. When Sam has left, his sister Lisa goes to the toy trunk. She likes to tease her brother. Lisa hides the rollerblades in the box! And then, she goes outside. Then, Sam comes back. He wants to rollerblade.'* Questions: *1) Where will Sam look for his rollerblades? 2) Why is Sam looking ... [there]? 3) Where does Sam think his rollerblades are? 4) Where are they really?*

Appendix B: Example pictures of the Theory of Mind Storybooks

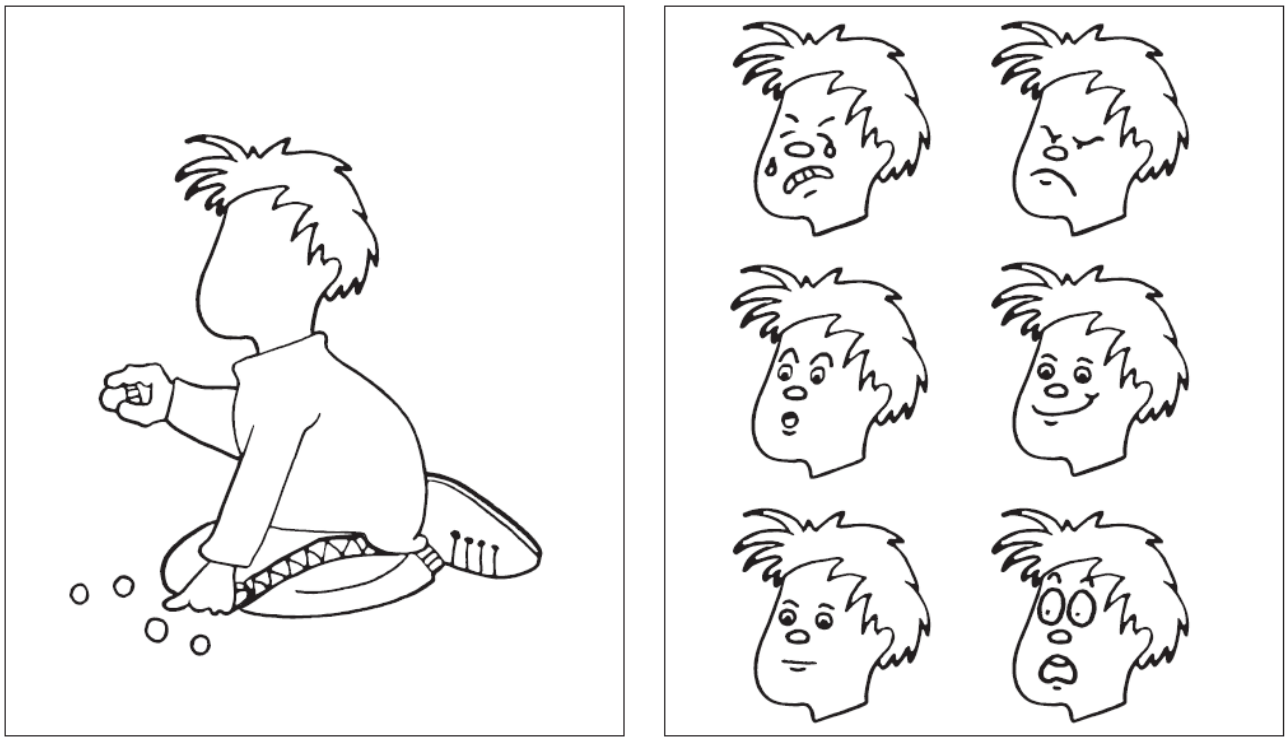

FIGURE 3. Emotion recognition task

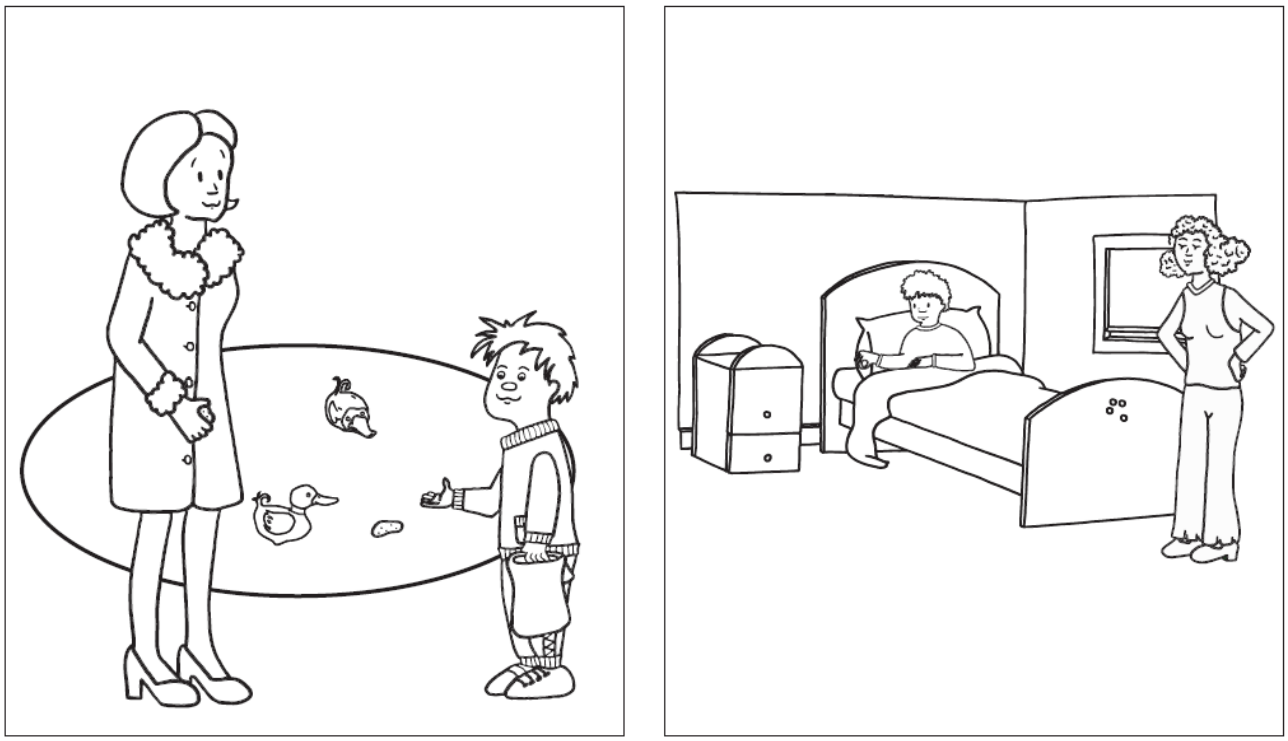

FIGURE 4. Mental-physical distinction task

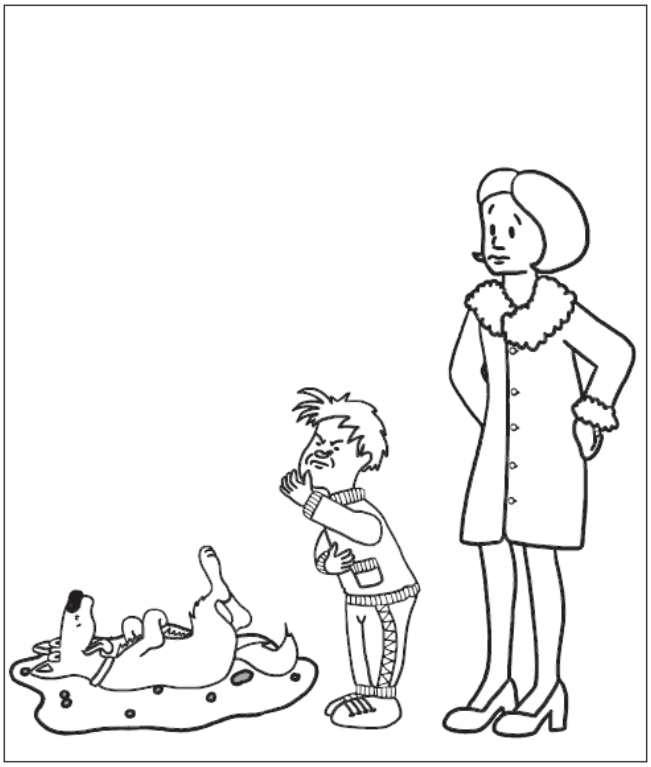

FIGURE 5. Close impostor task

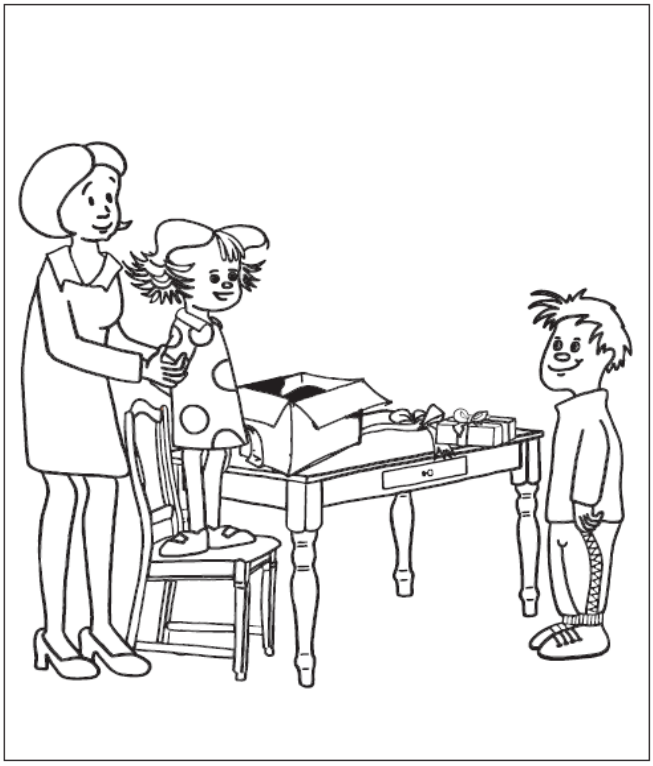

FIGURE 6. Seeing leads to knowing task

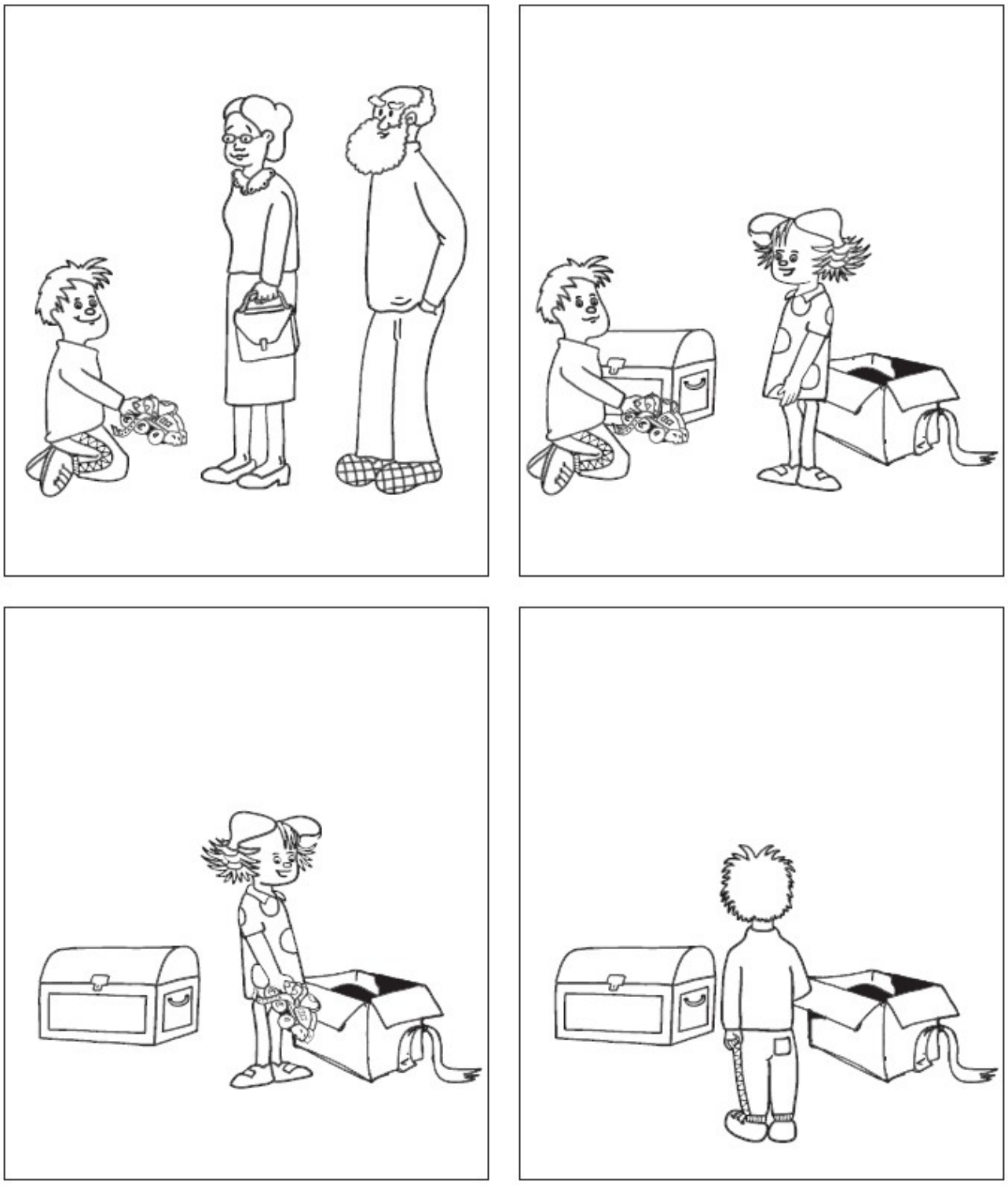

FIGURE 7. False belief task

Appendix C: Order of the tasks in the Theory of Mind Storybooks

Book	Task					Scoring of justification[3]	
	Nº	Name	Type	Quest.[1]	Max[2]	1 point	2 points
How is Sam feeling?	1	Emotion recognition	Happy	2 (1)	4	RM, GK & S	D, FB & VB
	2	Emotion recognition	Angry	2 (1)	4	RM, GK & S	D, FB & VB
	3	Emotion recognition	Scared	2	2		
	4	Emotion recognition	Sad	2	2		
	5	Emotion recognition	Surprised	2	2		
Sam goes to the park	6	Standard belief	Action	1 (1)	3	VRB	FB
	7	Standard belief	Emotion	2	2		
	8	Real-mental distinction	Pretend	4 (1)	6	LP	RR
	9	Desire	Action	1	1		
	10	Close impostor	Smell	4 (1)	6	IPP-almost	IPP & LP
	11	Desire	Emotion	2 (1)	4	VB, LP & S	D & RM
Sam goes swimming	12	Standard belief	Action	1	1		
	13	Standard belief	Emotion	2 (1)	4	LP, PC & S	FB & VB
	14	Desire	Action	1 (1)	3	RM & S	D
	15	Real-mental distinction	Dream	4 (1)	6	LP	RR
	16	Desire	Emotion	2	2		
	17	Real imaginary distinction	Think	4	4		
Sam visits his grandparents	18	Desire	Action	1	1		
	19	Explicit false belief	Action	2 (1)	4	VRB	FB
	20	Close impostor	Smoke	4 (1)	6	IPP-almost	IPP
	21	Not own belief	Action	1 (1)	3	VRB	FB
	22	Desire	Emotion	2	2		
	23	Real-mental distinction	Think	4 (1)	6	S	RR & LP
Sam at the farm	24	Standard belief	Action	1	1		
	25	Changed belief	Action	1 (1)	3	S	FB
	26	Real-mental distinction	Remember	4 (1)	6	LP	RR
	27	Not belief	Action	2 (1)	4	VRB	FB
	28	Desire	Emotion	2	2		
	29	Real imaginary distinction	Dream	4	4		
Sam's birthday	30	Perception knowledge	Know	1 (1)	3	LP	PC
	31	Desire	Emotion	2	2		
	32	Inferred belief control	Action	3	0		
	33	False belief	Action	3 (1)	5	LP & S	FB
	34	Inferred belief	Action	2	2		

Note. [1] number of test questions, and between brackets the number of additional justification questions; [2] maximum attainable points; [3] correct justification answers per task: D=desire, FB=fact belief, GK=general knowledge, IPP=insight physical process, LP=location possession, PC=perception criterion, RM=rest category mental state, RR=referring to reality, S=situational, VB=value belief, VRB=verb referring to a belief.