Validity & norming of the ToM Storybooks

E.M.A. Blijd-Hoogewys & P.L.C. van Geert (2004)

(See also Dutch publication: Blijd-Hoogewys et al, 2003)

Theory-of-Mind is an important condition for understanding the social environment and for showing socially adequate behavior (Astington & Jenkins, 1995). It is a skill that developes in children between their second and sixth year. Theory-of-Mind (which we shall abbreviate as ToM), refers to the ability to attribute mental states -such as beliefs, desires and emotions- to oneself and others and to use these mental states in understanding, predicting and explaining behavior of others and oneself (Mitchell, 1997; Premack & Woodruff, 1978). Desires are mental states referring to wishes, needs and demands; they are intentional and directed to uptaining something in the outside world. Beliefs are mental states such as thoughts and ideas; they are mental representations of reality, no direct copies of the world however. For instance, someone can think of ghosts and be afraid of them, but in fact they really do not exist.

ToM can be characterized differently at different ages, therefore, testing should involve a series of different tasks (Astington, 2001). In addition to providing a single, quantitative measure of the level of ToM knowledge, such a test has another advantage, namely that it allows us to measure all the components or aspects in the same child and thus to discover how these aspects are related during the course of development. Furthermore, it is important not to base the assessment on only one measurement (a so-called one-shot approach). A single assessment of the presence or absence of the ToM

knowledge of a child may be quick and efficient, but provides no information about the stability of ToM-knowledge. For long, test psychologists have recommended to implicate multiple tasks in multiple measurements, in order to reduce standard errors and make measurements more reliable and valid. Research has shown that such resulting compound scores are more stable, because they average over factors and lead to a more accurate measurement of the underlying skill (Hughes, Adlam, Happé, Jackson, Taylor & Caspi, 2000). In doing so, we might arrive at a more adequate diagnostic procedure, which can help us in studying the potential causes and nature of ToM differences (Hughes & Dunn, 1998).

In practice, such tests are seldom used. Quite the contrary, research is more often based on single measurements involving single aspects of ToM. To our knowledge there is only one test (also a Dutch test) that questions a wide variety of ToM-aspects: the ToM test (Muris and colleagues, 1999). This test however cannot be used in children younger than 5 years old, while the biggest development takes place at about four years (the mastering of false beliefs).

To conclude, there is a demand for a test that can estimate the general ToM knowledge of a child and that is based on multiple ToM components. For this reason, the ToM Storybooks were constructed. They address many aspects of ToM and its precursors (Van Geert, Hoogewys, Loth, & Serra, 1998), derived from established theoretical backgrounds. This new Dutch test (with at this moment also an Italian and Chinese version) was designed to measure the development and the mutual connection of different ToM-aspects over a broad age range. Since this is a new test, one should know its values

before using this test to compare different groups of children, for instance children with and without ToM-problems.

METHOD

Sample

The study includes 233 normally developing children (106 girls and 127 boys). The data were obtained from several research projects, conducted over a period of four years. Ages range from three to eleven years, with a main focus on children from three to seven years. Fewer children were tested after this age (see table 1 for a distribution of the different ages).

All children came from kindergartens and elementary schools in the province and/or city of Groningen. All children have a Dutch linguistic background, and did not have language acquisition problems that could have hampered their performance on ToM tasks (see for instance Garfield, Peterson & Perry, 2001; Lohmann & Tomasello, 2003).

Instrument

The ToM Storybooks (a revision of the test used in Serra, Loth, van Geert, Hurkens & Minderaa, 2002) is a test that measures a variety of ToM-components: emotion recognition tasks, tasks questioning the difference between physical and mental entities, tasks on the understanding that seeing leads to knowing, desire tasks and belief tasks (like false belief tasks) (for more detailed information about the test, see Blijd-Hoogewys & van Geert, submitted).

All tasks are presented in the context of a story and illustrated by full color pictures, also using caressable furs, toy doors that can be opened, and magnetized emotion faces that can be placed on the figures. The administering of the test takes 40 to 50 minutes, including a short break. There are six books, resulting in a maximum total score of 112 (quantitative & qualitative answers).

There are four versions of the ToM Story Books: Sam, Lotje, Pieter, and Hanna, all consisting of six storybooks. Each version has a different protagonist and different stories, but is based upon the same underlying test structure. These alternative versions can be used in a longitudinal design, preventing trivial learning effects that might result from mere repetition.

RESULTS

Study 1: Norming the ToM Storybooks

Method

Calculating norms for the ToM Storybooks: preceding remarks

Before making standard norms, we first checked if it was necessary to calculate separate norms for boys and girls; because on average, girls have slightly higher ToM total scores on the ToM Storybooks than boys (53.77 versus 51.63 respectively). An independent samples T test showed no significant differences in their performance (two sided, $p=.188$). Also, the variance within each group could be considered equal (Levene's test, $p=.650$). Therefore, norms for girls and boys together were considered sufficient.

Making norms

Traditionally, the expected total score determines a norm for a specific age. It is based on the distribution of all ToM total scores at that age summed up in one number, namely the sum of all scores multiplied by the probability that this score will occur at that age (Traub, 1994). Because one will never possess all possible scores at any specific age, the making of norms is customarily based upon age groups. In that case, the expected score can be estimated by calculating the average score for each age group. However, this method still requires a substantial number of children tested at each age group. Since our total group consisted of only 233 children spread over a wide age range, the use of such age groups is hardly possible.

ToM total scores normed

For this instrument, norms for the total scores were obtained by using a quadratic conversion curve based upon a locally weighted moving regression (with a window of 40%). In studies on norming, conversion curves are frequently used to determine norms (see, for example, Snijders, Tellegen & Laros, 1988).

Loess smoothing

For every observed age, a neighboring age group was determined with members left and right from this chosen age (for instance for a chosen age of 80 months we chose a neighboring group of children ranging from 46 to 84 months). Next, we estimated the expected score, for instance for the age of 80 months, by fitting a quadratic curve over the scores of this age group. Because of the large sample of children (233), the borders of

every matching age group needed to be taken wide enough to achieve a reliable estimate of the quadratic model. We implemented these requirements in the form of a locally weighted moving regression [1]. The Loess procedure computes the quadratic regression model of the first 40% of the raw data, taking into account their individual weights - meaning that the scores in the central part of the window have bigger weights than those on the extremes. For instance, when we calculate the expected score for an age of 80 months the observed score of a 79-months-old has a larger weight in this estimation procedure than the score of a 76-months-old. The average point is calculated for 0-40% of the data; this is repeated for the next 1-41%, 2-42%, 3-43% and so on. Finally, all the calculated average points are combined into a smooth curve of expected scores.

This Loess smoothed curve is compared with the curve based on a simple continuous growth model, with the following form:

$$y = a + b / x^2 + \text{error} \quad \text{(equation 1)}$$

and yields an r^2 of 0.59 (a = 9.679; b = - 5094.03).

A B-spline interpolation

Since the raw data contain fewer observations in the region of the higher ages, the Loess smoothing results in an unbalanced curve. To correct for this, a B-spline interpolation of the Loess-smoothed data was calculated. A B-spline is a continuous curve that connects all points of the Loess curve (calculated scores separated by equal time intervals, thus

[1] The Loess or Lowess (a locally weighted least squares estimate) smoothing procedure is a fitting technique that follows the local distribution of the data as reliably as possible is.

equal age distances).[2] The result is a continuous model that provides a good fit of both the Loess-smoothed data and of the raw data (r^2=0.98 and r^2=0.60 respectively).

The Loess smoothed curve is used to calculate the expected score for every age in the age range. Next, the difference between the expected score and the observed score at all ages are calculated. The square of those differences produces the observed variance for every age in the age range. Then, we can use the same Loess technique (with the same window size) to estimate the expected variance at every age. The square root of this expected variance results in the expected standard deviation for that age. Because we described the Loess curve on the basis of a B-spline, we can calculate the expected score and expected standard deviation for every arbitrary age. These scores form the basis of our norm score.

Z-score

Imagine an individual with an age A has a score of S_0. We convert this S_0 score in a z-score. In order to do so, we search in the Loess model (based upon age L) for the estimated score S_M (model score) and the estimated standard deviation D_M. Z-scores reflect how many standard deviations (D_M) a score is away from the average test score (S_M). The z-score is defined as:

$$Z0 = (S_0 - S_M)/ D_M$$

ToM-Q & age equivalent

[2] The estimation of the curve is carried out with the aid of the SPSS program Table Curve 2D; this program matches the Loess procedure to the calculation of a B-spline.

As a last step in the norming of the ToM Story Books, the ToM Quotient (ToMQ) was computed. The ToMQ is a quotient with an average of 100 and a standard deviation of 15. The minimum was set at 55 and the maximum at 145 (average of 100 minus or plus three times the standard deviation. A ToMQ of 100 is based upon a z-score of zero thus 55 + (3 + 0)*15 = 100. The formula to calculate the ToMQ is:

ToMQ = 55 + (3 + z-score)*15

In practice, the calculation of ToM-Q's is carried out with the help of an Excel file in which two ToMQ calculation functions are defined, one on the basis of only the quantitative data and the other on the basis of both quantitative and qualitative data (respectively ToMQ-Q en ToMQ).[3] In addition, we also wrote an Excel function to calculate the age equivalent of a child, i.e. the age for which the expected score (equal to a ToM of 100) equals the child's observed score.

An advantage of a quotient is that it provides a simple representation of good and bad scores. For instance, if a child has a ToMQ of 102, this is considered to be an average ToMQ. If a child has a ToMQ of 63, this is considered to be deviant. A ToMQ is deviant if a difference of two standard deviations below a ToMQ of 100 occurs, that is, a ToMQ lower than 70.

In some cases, children have total scores that are so low that a ToMQ cannot be calculated. That is, the score falls beyond the reach of the norm table, which is a ToMQ lower than 55 (three standard deviations below average). In this case, an age equivalent

[3] A quick-and-dirty estimation of the ToM functioning of a child can be made on the basis of only the quantitative data. Taking the qualitative answers into consideration takes far more time since they consist of open answers not easy to judge. However, especially in children with ToM problems the qualitative data can give more insight in the ToM knowledge underlying the child's score.

can be calculated. However, it needs to be interpreted with extreme caution, since it only gives an estimate of the age level upon which a child functions on ToM.

Subscores normed

The ToM Storybooks specify seven subscores, based upon theoretically derived aspects within ToM: emotion recognition (ER), mental physical distinctions (MPh), real imaginary distinctions (RI), close impostors (CI), desires (D), beliefs (B, without FB) and false beliefs (FB). [4]

Percentiles

Because of the relatively small number of items in each subscore (minimum of 9 and maximum of 34), the norming of the subscores is based upon percentiles. For each age, an estimation of percentiles for every subscore needs to be calculated.

The estimation of percentile borders is based upon moving averages with a window of 15. In order to use moving averages, the observed scores are first ordered by age. Next, the 5^{th}, 10^{th}, 20^{th} and so on until the 100^{th} percentile are calculated for the first fifteen scores. Then, the same is done for the second till the sixteenth score, the third till the seventeenth, and so on. In the end, the averages of these calculated percentiles are presented as moving percentile curves. After that, these moving percentile curves are smoothed with a Loess procedure resulting in continuous percentile lines for the total age range. By way of example, figure 6 shows the moving percentiles for false beliefs. The lowest, the middle and the upper bold line represent percentiles 5, 50 and 95 respectively.

[4] 'Seeing leads to knowing' does not result in a separate subscore, since only three questions are asked within this area.

[Figure 6: Percentile landscape for false beliefs]

An Excel function was written for determining the percentile border of a specific subscore for a specific age and a specific score. We see that the scores increase with age; however, some of the oldest children are still unable to accomplish the tasks. A skewed distribution is found in both the youngest group and the oldest groups, showing that in young children lower scores occur more frequently than higher scores. Only a few children have exceptionally high scores. The opposite picture is true for the oldest group, with more children succeeding on the test and a few with exceptionally low scores.

Study 2: Validity of the ToM Storybooks

Method

Validity of the ToM Storybooks

The operationalization of ToM is theoretically justified since the test is build on relevant ToM aspects as known from the literature (see Blijd-Hoogewys & van Geert, in press).

Analyses show that the ToM scores are positively correlated with *chronological age* ($r = .67$; $p < 0.001$) (compare Wellman et al, 2001). The *internal consistency* of the ToM Storybooks is high ($\alpha=.91$), even after correction for the influence of age (n=218, age=3-8, Cronbach's $\alpha=.81$) (Volkers, 2002). This is consistent with findings from

comparable research on standard and complex false belief tasks (Cronbach's α=.84; Hughes e.a., 2000) and suggests that the different tasks measure the same construct.

The *test-retest reliability* is also satisfying (n=49, age=3-7, r=.74, p=.001; when corrected for age r=.87, p=.001) and consistent with findings from comparable research (r=.77: Hughes e.a., 2000). This is not lower than test-retest correlations on most standard psychometric measures on cognitive skills which have a greater number of items. However, it needs to be noted that the scores rose significantly when the test was administered twice in a short time span (p<.001) (de Groot & van der Honing, 2001). A similar observation has been reported by Muris and colleagues (1999). Nevertheless, the increase in ToM total scores found in our research was never higher than half a standard deviation. Such a rise should not be considered an unwanted effect, since it can be expected that young children in particular learn from being tested.

The effect of different *test administrators* on ToM scores was studied in a smaller research project (n=27, age=3-7). Half of the subjects were assigned to one test administrator while the others were assigned to a second one. The results showed no effect of test administrator (de Groot & van der Honing, 2001). The *inter-rater reliability* based upon the results for the qualitative answers was high (Cohen's Kappa = .81 to .97 for the 21 categories, .97 to .99 for the 0-2 point scores) (n=10 control children and n=10 children with PDD-NOS, age=3-11, N=5 raters) (Boeting & Wolters, 2003).

In order to determine the *reliability of the parallel forms*, correlations, a classic measure of stability, were calculated. This was only done for the total scores. Taking into account the binomial distribution of the subscores, item response theory predicts a certain level of variation and as a result correlation will be suppressed. In light of this expected

variation, the correlations found between three consecutive measurements within a period of maximally a month, were high (Sam1 & Lotje, r=0.91, Sam1 & Sam2, r=0.86, and Lotje & Sam2, r=0.90) (de Groot & van der Honing, 2001). It can thus be concluded that the reliability of the parallel forms is satisfying

As reported before, children have a free choice of the order of four books. This *different order* poses no problem for later analyses of the data, since the four books are alternative versions with an identical underlying structure. However, analyses have shown that they do differ in degree of difficulty (p=.01). As a result, the chosen order of books might influence the test results, for instance as a consequence of differential fatigue or boredom. In that respect, it should also be taken into account that the test takes about 45 minutes. Eventual effects of fatigue or boredom were calculated in the following way. First, average scores of each storybook were calculated (not including book 1 and 6, because they always occur in the first and last position respectively). By means of a permutation test for dependent measures, p-values were calculated for differences between the storybooks. However, since the children choose the order of four books themselves (namely books 2 to 5), we need to check whether the mixture is sufficiently random or not (it is possible for instance that almost all children choose book 5 first, because it has the most attractive title, for instance). The actual distribution of the books is then used to calculate an expected score for each step in the test procedure. The latter score is calculated by multiplying the number of specific books with the average score of the book in question and dividing it by the total number of books administered.

[Table 2: average scores based upon positions of the books read]

	position 1	position 2	position 3	position 4
measurement 1				
observed average	0.553	0.523	0.518	0.522
expected average	0.568	0.547	0.565	0.558
measurement 2				
observed average	0.572	0.578	0.560	0.596
expected average	0.555	0.545	0.567	0.572
measurement 3				
observed average	0.561	0.557	0.582	0.594
expected average	0.560	0.549	0.556	0.575

Table 2 compares the scores for each position (rank order) for the expected scores and the observed scores (these data are derived from the research on the reliability of forms, in which 3 measurements were used: version Sam, version Lotje and version Sam again). Only measurement 1 (Sam 1) shows a decline in the average score towards the end of the testing procedure; it consists of 1/5th of a standard deviation of the actual scores. The expected decline on the basis of the average score of the storybooks is 0.01. The observed decline is 0.03, i.e. 0.02 bigger than the expected decline. However, this decline is not statistically significant and there is no evidence for a negative effect of increasing fatigue or boredom on the test scores. It can be concluded that the different conditions (different orders) of testing did not affect the results in a meaningful way.

Since ToM questions make a relatively strong appeal to lexical and syntactic knowledge, we expect to find a positive *relationship between ToM scores and scores on a language test* (see for instance Garfield, Peterson & Perry, 2001; Lohmann & Tomasello, 2003; Serra, 1998). However, since ToM questions address a particular kind of knowledge, namely knowledge about the mind, we expect that the positive association is not very strong. In a smaller study with 118 children between 3 and 7.5 years old, we found a low correlation between the ToM Storybooks and language acquisition tests (r=.27) (Huyghen, 2000). This implies that about eight percent of the variance in ToM total scores can be explained on the basis of language acquisition scores.

If our test provides a valid measurement of ToM, it should show substantial *correlations with other instruments* and tests that focus on the child's knowledge of the mind or that measure skills directly related to such knowledge. A test that measures such knowledge is the CSBQ (Children's Social Behavior Questionnaire; Luteijn, Luteijn, Jackson, Volkmar & Minderaa, 2000; Dutch version: VISK; Luteijn, Minderaa & Jackson, 2002), in particular the subscales 'orientation problems' and 'not understanding'. The CSBQ is a standardized instrument that describes autism-related behavior of children. The subscale 'orientation problems' refers to the inadequate automatic orientation of oneself in time, place, activity or person; the subscale 'not understanding' refers to difficulties in understanding and perceiving social information. As predicted, the correlations of the ToM total score with these ToM related subscales were negative (n=39 children with PDD-NOS and/or ADHD, 3.5-11.5 years old, r=-.48 and r=-.49 respectively, p=.01, there is a negative correlation since a lower score implies a higher level of problems) (van Pagée, 2002). The lower children score on the ToM Storybooks, the more problems they exhibit on the CSBQ-subscales. A comparable correlation was found between the ToM Storybooks and the VABS questionnaire subscore Interactive Sociability (n=43 children with PDD-NOS and/or ADHD, age=3-11, r=.45, p=.01; n=109 control children, age=3-8, r=.30, p=.01; a positive correlation since a higher score implies higher sociability) (Holwerda, 2003). The VABS questionnaire consists of theoretically derived items on active and interactive sociability (Frith, Happé & Siddons, 1994; Dutch translation: Hoogewys, van Geert & Serra, 1999). The first can be performed without the ability to mentalize, whereas the second cannot be performed without a ToM.

A test that measures ToM should also be able to *distinguish children with ToM problems* from those who show a normal ToM development. Research showed that children with PDD-NOS have significantly lower ToM total scores than normal children on the ToM Storybooks (n=38 versus n=233, p=.01). Also children with ADHD and children from a lower social background had significantly lower scores on the ToM Storybooks (n=15, p=.01 and n=31, p=.05 respectively) (Blijd-Hoogewys, Serra, van Geert & Minderaa, 2002). Children with PDD-NOS showed a specific deficit in understanding and predicting emotions. They had significantly lower scores on the emotion recognition tasks, the desire-emotion tasks and the belief-emotion tasks. They also performed worse on the real-imaginary tasks (n=11 children with PDD-NOS versus n=23 normally developing children, p=.05). They gave less ToM-related answers and made more mistakes ($p<.05$ and $p<.01$, respectively). They also referred less to beliefs in their answers ($p<.05$). No differences were found in terms of desires. To conclude, they not only showed a delay (a slower quantitative growth) but also a possible deviance (a different qualitative pattern of change) in their ToM development (Serra et al, 2002).

CONCLUSION & DISCUSSION

It can be concluded that the validity of the ToM Storybooks complies with the requirements made to an instrument of this sort and also that the norms are adequate and practically usable. As a consequence, this measurement instrument may have potential for a range of applications to both fundamental and applied work.

Since no differences were found between boys and girls, common norm tables were considered sufficient. The norm score is referred to as the ToM quotient or the ToMQ.

As regards the validity of this new ToM test. the internal consistency is high, the test-retest reliability is good, the effect of different test administrators is low and the inter-rater reliability of open answers is high. Also, the reliability of parallel forms is satisfying and there is no effect of fatigue. The latter findings are consistent with those of Wellman and colleagues on false belief tasks (2001) 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.

The ToM Story Books have an expected low but statistically significant correlation with language acquisition tests and substantial correlations with other instruments testing the knowledge of the mind, e.g. the CSBQ and the VABS questionnaire (subscore interactive sociability). The test has good discriminative power. It can be used to distinguish children with a normal ToM-development from children with ToM problems, like for instance children with PDD-NOS, ADHD or children from a lower social background.

Limitations and potentialities of the ToM Storybooks

One of the restrictions of this research is that it had fewer children in the older age regions, which implies a reduction in reliability. This is due to the fact that the test is primarily intended for younger children, up to six or seven years old. We used a

conversion curve that took into account the skewed age distribution. Still, the norms for the older children should be taken with caution. To be more accurate, additional tasks should have been included for testing these older children, like for instance second-order beliefs tasks (see for instance Hughes et al, 2000). We have now added such tasks to the ToM Storybooks (in the form of an additional reading book) (op het Veld & Van Royen, 2003) and are currently undertaking research with these supplements.

There was a small learning effect in comparing the test-retest correlations. 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 posttest can eventually be considered a sound predictor of learning abilities. The eventual absence of a comparable rise in specific groups of children, for instance children with autism, could provide interesting information about the nature of ToM abilities in such children. In this line, further research on ToM might profit even more from dynamic testing ideas – as opposed to static testing – where the learning potential of a child is quantified on the basis of its understanding and use of feedback given during testing. This takes us back to Vygotsky's zone of proximal development, where the focus is not on what has already developed but what is in the process of developing. Also Binet, the founder of static testing, advocated such a process assessment (in Grigorenko & Sternberg, 1998).

LITERATURE

Blijd-Hoogewys, E.M.A., Huyghen, A.N., van Geert, P.L.C., Serra, M., Loth, F.L., & Minderaa, R.B. (2003). Het ToM Takenboek: constructie en normering van een

instrument voor het meten van 'theory of mind' bij jonge kinderen [The ToM Story Books: construction and norming of an instrument measuring 'theory of mind' in young children]. *Nederlands Tijdschrift voor de Psychologie*, 58, 19-33.

Blijd-Hoogewys, E.M.A., Serra, M., van Geert, P.L.C., & Minderaa, R.B. (2002). Theory of Mind: denken over denken, willen en voelen. De ontwikkeling van een nieuwe test voor kinderen [Theory of Mind: thinking about thinking, wanting and feeling. The development of a new test for children]. *Wetenschappelijk Tijdschrift Autisme*, *1*, 4-13.

Blijd-Hoogewys, E.M.A., & van Geert, P.L.C. (submitted). The developmental pattern of Theory-of-Mind knowledge, potential discontinutiy and inter-individual differences.

Boeting, B.A., & Wolters, C.W. (2003). Hoe komt het dat …? Een k*walitatieve analyse van het ToM Takenboek* [How come …? A qualitative analysis of the ToM Story Books]. Unpublished master's thesis, University of Groningen, Groningen, the Netherlands.

De Groot, A.C., & van der Honing, N.M. (2001). *Als je begrijpt wat ik bedoel. Validering en betrouwbaarheid van het ToM Takenboek* [If you know what I mean. Validation and reliability of the ToM Story Books]. Unpublished master's thesis, University of Groningen, Groningen, the Netherlands.

Frith, U., Happé, F., & Siddons, F. (1994). Autism and theory of mind in everyday life. *Social development*, 3, 2, 108-124.

Garfield, J.L., Peterson, C.C., & Perry, T. (2001). Social Cognition, Language Acquisition and The Development of the Theory of Mind. *Mind & Language*, 16 (5), 494-541.

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

Holwerda, F. (2003). De psychometrische kwaliteiten van de VABS-v. ToM-vaardigheden van het kind via de ouders gemeten [The psychometric qualities of the VABS-questionnaire. ToM abilities of children indirectly measured in parents]. Unpublished master's thesis, University of Groningen, Groningen, the Netherlands.

Hoogewys, E.M.A., van Geert, P.L.C., & Serra, M. (1999). *Dutch translation of the Vineland Adaptive Behavior Scales: supplementary items.* Groningen: Rijksuniversiteit Groningen.

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

Huyghen, A.M.N. (2000). *Denken overdacht. Normering van het ToM Takenboek* [Thinking rethought. The norming of the ToM Story Books]. Unpublished master's thesis, University of Groningen, Groningen, the Netherlands.

Lohmann, H., & Tomasello, M. (2003). The Role of Language in the Development of False Belief Understanding: A Training Study. *Child Development*, 74, 1130-1144.

Luteijn, E.F., Luteijn, F., Jackson, A.E., Volkmar, F.R., & Minderaa, R.B. (2000). The Children's Social Behavior Questionnaire for milder variants of PDD problems: Evaluation of the Psychometric characteristics. Journal of Autism and Developmental Disorders, 30, 317-330.

Luteijn, E.F., Minderaa, R.B., & Jackson, A.E. (2002). *Vragenlijst voor Inventarisatie van Sociaal gedrag bij Kinderen, Handleiding* [The Children's Social Behavior Questionnaire, Manual]. Lisse: Swets & Zeitlinger.

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 (1), 67-80.

Op het Veld , A.M., & van Royen, P. (in press). *Onderzoek naar de Second Order Beliefs bij kinderen met PDD-NOS* [Research on Second Order Belief tasks in children with PDD-NOS]. Unpublished master's thesis, University of Groningen, Groningen, the Netherlands.

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.

Snijders, J. Th., Tellegen, P. J., & Laros, J. A. (1988). Snijders-Oomen niet-verbale intelligentietest SON-R 5½-17. Verantwoording en handleiding. [Snijders-Oomen non verbal intelligence test SON-R 5½-17. Justification and manual.] Groningen: Wolters-Noordhoff.

Traub, R.E. (1994). *Reliability for the social sciences: theory and applications*. Thousand Oaks: Sage.

Van Geert, P.L.C., Hoogewys, E.M.A., Loth, F.L., & Serra, M. (1998). *ToM Takenboek* (testmateriaal) [ToM Story Books]. Groningen: Rijksuniversiteit Groningen.

Van Pagée, M. (2002). *Waar zal Sam gaan zoeken? Theory of mind-vaardigheden bij kinderen met PDD-NOS en kinderen met ADHD* [Where will Sam look? Theory of Mind abilities in children with PDD-NOS and children with ADHD]. Unpublished master's thesis, University of Groningen, Groningen, the Netherlands.

Volkers, J.H. (2002). *Een nadere analyse van het ToM Takenboek: item-analyse en betrouwbaarheid* [A closer analysis on the ToM Story Books: item-analysis and reliability]. Unpublished master's thesis, University of Groningen, Groningen, the Netherlands.

Wellman, H.M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development*, *72*, 655-648.