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

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

CHAPTER 3

Norming the ToM Storybooks
A comparison between two methods

Abstract: Although false beliefs tests are valuable for scientific research on Theory-of-Mind (ToM), clinical and applied use requires more comprehensive tests, containing multiple tasks on different aspects of ToM. In order to compare children, normalized scores for these tests are also necessary. The ToM Storybooks is a comprehensive ToM test focussing on basic ToM components children develop between their third and sixth year of life. The test can be administered in children with (or without) ToM problems up to the age of twelve and results in an assessment of the general ToM knowledge of a child. In order to calculate norm scores for this test, 324 typically developing children (3-12 years old) were tested. The ToM sumscore was normed in two distinct ways: based on a monotonically rising regression model and based on a Loess (locally weighted least squares estimate) smoothing procedure. Because gender differences were obvious, norms for boys and girls separately were computed. Based on z-scores, ToM quotient scores were calculated, in addition to confidence intervals and age equivalents. Both norming methods were compared. The Loess smoothing procedure, which more faithfully follows plateaus and temporary developmental regressions in the raw scores, was judged to be superior.

This chapter is based on: Blijd-Hoogewys, E.M.A., Huyghen, A.N., Geert, P.L.C. van, Serra, M, Loth, F., & Minderaa, R.B. (2003). Denken overdacht. De normering van het Theory-of-Mind Takenboek. *Nederlands Tijdschrift voor Psychologie.*

INTRODUCTION

Theory-of-Mind (abbreviated as ToM) is the social cognitive ability to attribute mental states to oneself and to others and to use these in understanding, predicting and explaining behaviour of oneself and others (Mitchell, 1997; Premack & Woodruff, 1978). ToM thus relates to the ability to adequately take into account the desires, beliefs and feelings of others. ToM research is typically based on the paradigm of false beliefs, often tested using the Sally and Anne test (from Baron-Cohen et al., 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. Children are then asked where Sally will look for her ball when she returns. Children who answer 'in the basket' are considered to understand false beliefs. Children who answer 'in the box' do not (yet) comprehend false beliefs and are misled by their own true belief.

Beginning in the mid 80's, a considerable amount of research has been undertaken on ToM, both in the field of developmental psychology looking into the normal development of ToM (for a meta-analysis see Wellman et al., 2001) and in the field of clinical psychology, focusing on the delayed and deviant development of ToM (e.g. Baron-Cohen, 1989b, 2000; Corcoran et al., 1997; Peterson et al., 2005; Yirmiya et al., 1998). This research has addressed the question of the age at which ToM develops, which (clinical) groups show ToM-problems and which factors influence ToM functioning such as language, executive functions and intelligence (for a review see Baron-Cohen et al., 2000). Most research has focussed on false beliefs, though ToM comprises of far more, like the understanding of desires and emotions.

Studies on the normal development of ToM show that it develops during the first six years. It evolves from a simple desire theory (only taking into account desires) to a complete belief-desire theory, from true beliefs to false beliefs, and from the understanding of first-order beliefs (thinking about the thoughts of another person) to second-order beliefs (thinking about the thoughts someone has about the thoughts of a third person) (Wellman, 1990). In the end, children comprehend that mental processes are subjective and independent of reality. They understand that different people can interpret the same event differently and that these interpretations can also result in different emotions.

Testing ToM in applied and clinical settings

In applied and clinical settings, ToM has become increasingly important. In the treatment of autism, for instance, ToM-problems are seen as an important cognitive explanation, next to a weak central coherence and problems with executive function (for guidelines in using these theories in clinical practise, see for instance Williams & Wright, 2004). A major problem in clinical settings is how to accurately determine ToM-problems in individual children. Are the tests used in scientific research applicable to clinical practice?

For tests to be used in an applied setting, they need to comply with a number of requirements; the most important of which are the following. First, the test should be easily administrable. Second, children with ToM-problems should be easy to detect. Third, the test must be able to specify the magnitude of ToM-problems in a particular child.

Most tests used in scientific research fulfil the first two requirements. Using the Sally and Anne test as an example; the test is easy to administer, since one only needs two dolls, a blanket, a box, a basket and a ball. The instructions are simple and easy to translate from the English version. In addition, age limits for the understanding of false beliefs have been found consistently over different tasks, different executions and different cultures (Callaghan et al., 2005; Wellman et al., 2001; Wellman & Liu, 2004), so children with FB problems can be detected. For instance, young children with autism fail on these tests (Yirmiya et al., 1998).

The third requirement, relating to the discriminative power of the test, is less often achieved since standard first-order false belief task are not sensitive enough to detect children with less profound ToM-problems, such as children who use compensatory strategies (for instance with the help of alternative, linguistically based, routes; see Fisher et al., 2005). These children do succeed on ‘simple’ false beliefs tests. Next to that, such ToM tasks often contain only a few questions (mostly two to three); consequently, these tests do not differentiate much (for a critical review on the use of false belief tasks, see Bloom & German, 2000).

In order to capture more subtle ToM problems, a comprehensive test is required (Blijd-Hoogewys et al., 2008; Hughes et al., 2000). Since ToM is a multi-dimensional construct, not only consisting of false beliefs, such a test should incorporate tasks addressing a wide variety of mentalizing skills. It should also lead to adequate discrimination between development in different ToM aspects. A comprehensive test results in a wider range of possible total scores, enabling variation in scores, and it leads to better

differentiation between children. Examples of comprehensive tests are the ToM battery and Strange stories battery of Happé (1994), the Tom-Test of Steerneman and colleagues (2002), the advanced test of ToM of Kaland and colleagues (2002), the ToM Tasks of Tager-Flusberg (2003), and the ToM Tasks of Wellman and Liu (2004).

As regards the design of comprehensive tests in general, one can distinguish two different types: a staircase design and a complete design. For the first type, a child has to succeed on a preceding module in order to be presented with the next module. These tests adhere to the idea of successive progressively developing abilities, postulating a consistent, almost monotonic development, where a child cannot comprehend certain tasks before it masters its precursors. The second type, using a complete design, does not postulate such a development and therefore all tasks are administered to all children, irrespective of their age or ability.

The existing comprehensive tests used in ToM research appear to prefer the staircase design. For instance, in Tager-Flusberg ToM test (2003) children are presented with tasks on 'perception/knowledge and first order false beliefs' only if they first succeed on the 'pretend and desires' tasks. The staircase design used in ToM research coheres well with the developmental orders found in normal children, as shown in the meta-analysis of Wellman and Liu (2004) which compared more than 170 studies. They concluded that the developmental order is not one of addition or substitution, but one of modification or mediation (pp. 536). Initial ToM understanding is broadened or generalised to more mature insights: from the understanding of desires, over knowledge and ignorance, to false beliefs and then to hidden emotions.

But, children with ToM problems need not follow the same developmental order. It is not sure if children with autism have a delay or a deviance in their ToM development (Serra et al., 2002). As Baron-Cohen (1991b) showed, children with autism are not only delayed in mastering ToM-tasks, but they also acquire these skills in a different order than normal children and even than children with a mental retardation. Recent studies have affirmed this atypical developmental order (Peterson et al., 2005). As a result, the staircase development found in normal children may not apply to certain clinical groups of children. Therefore, we recommend the use of comprehensive ToM tests based on a complete set of ToM skills instead of staircase-designed tests.

In order to compare children and evaluate their inter-individual differences, normative scores are required. At the moment, only a few

comprehensive ToM tests have been developed and even fewer of these instruments have been normed. An exception is the ToM-Test (Muris et al., 1999; Steerneman et al., 2002), which uses percentile norms and is widely used in applied and clinical settings in the Netherlands. This test has norms for children from five to twelve years old. However, it is known that five-year-olds already have development many ToM-concepts (e.g. Wellman, 1990). To be able to test children from three years on, we developed a new test - the ToM Storybooks - that focuses on basic ToM components.

The ToM Storybooks

In this article, we present the norming of the ToM Storybooks (for the construction, reliability and validity of the ToM Storybooks see Blijd-Hoogewys et al., 2008). This test was developed to obtain an assessment of the basic ToM knowledge of a child. The test can be used in children from three years upwards. It involves tasks which typically developing five year-olds should be able to master. However, since children with ToM problems are hypothesized to have an underdeveloped ToM (their ToM is not maximally developed between three and six), we decided to construct norms for a broader age range, from 3 to 12 years old.

Different norming approaches

In order to calculate normative scores, one can use different approaches. Raw scores can be converted to percentiles, deciles, quintiles, quartiles, Z-scores and quotient scores. For the ToM Storybooks scores, we have chosen to calculate quotient scores over other methods. Quotient scores are probably the best known type of normed scores because they are used in intelligence tests (IQ scores). They have an average of 100 and a standard deviation of 15 and they provide a simple representation of high and low scores, which are easily understood by practitioners.

A norm can be defined as the expected score determined for a specific age. It is based on the distribution of all 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 based upon age samples and the differences between age groups are smoothed. In the past the latter was done manually by using French curves (Angoff, 1971;

in Zachary & Gorsuch, 1985). Nowadays, continuous norming is performed using statistical smoothing (Zachary & Gorsuch, 1985; Taylor, 1998). In that light, different conversion curves can be used. Determining which one should be preferred is not simple, since they are based on different theoretical assumptions. This article presents two norming methods: one based on a monotonically rising regression model and one based on a Loess smoothing procedure. Both result in a quotient score, called the ToM-Q.

Norming based on a monotonically rising regression model

The first norming method originates from methods used in intelligence norming. Norms are calculated on the basis of curve fitting using regression analysis.

This method is based on two assumptions: 1) there is a continuous positive change in the ToM sumscore and 2) this rise is monotonic. Temporal developmental regressions, sudden jumps or plateaus do not comply with these assumptions, except for a plateau at the end of the age period when the measured construct is considered to be consolidated.

Norming based on a Loess curve

The second norming method uses a different smoothing technique. It is based on a Loess or Lowes, a locally weighted least squares estimate. Smoothness of the estimated curve is induced by weighing neighbouring observed scores (see for instance Simonoff, 1996; Härdle, 1991). The method assumes continuous change. However, unlike the former method, the change need not be monotonic. Temporary plateaus and developmental regressions are allowed, if they are empirically obvious. This is the primary reason for using a Loess curve fit, since plateaus and developmental regressions are well-documented phenomena in developmental psychology (for a general framework of developmental spurts and plateaus see Fischer & Bidell, 2006; for a discussion of regressions in development see Siegler, 2004).

The Loess procedure evaluates consecutive windows of data. It first calculates moving regressions and then computes a quadratic regression model of the first subset of the raw data, taking into account their individual weights. The scores in the central part of the window are assigned larger weights than those on the extremes. We have chosen to calculate the first average point for 0-30% of the data; this is repeated for the next 1-31%, 2-32%, 3-33% and so on. Finally, all the calculated average points are combined into a smooth curve of expected scores.

In our view, the 30% window size yields a good compromise between a maximally faithful Loess procedure that follows the data quite closely (too small windows result in a capricious curve, similar to just displaying the raw data) and a simple model fit (with the largest window possible, that is 100%, similar to a quadratic model). The 30% window results in a curve that is continuous, but also follows the local deviations.

Norming the ToM Storybooks

In this article, we present two different ways of norming the ToM Storybooks. First, we calculated standardized scores based on the assumption that a measurement norm for a growth process must be a monotonically rising function, irrespective of whether actual subjects can show developmental regressions or plateaus. Second, we calculated standardized scores based on the assumption that developmental regressions and other anomalies are not only allowed but are in fact expected in the empirical data. Depending on one's theoretical preference, one method is more suitable than the other. However, the key questions should be: Which norming curve leads to the best description of the population? And how can I it be used in clinical practice?

METHOD

Subjects and setting

The children came from preschools, kindergartens and elementary schools, from both provincial and urban regions in the Netherlands. All children had a Dutch linguistic background, and did not have any language acquisition problems that could have hampered their performance on the ToM tasks (for the effect of language on ToM performance see for instance Astington & Baird, 2004; Garfield et al., 2001; Lohmann & Tomasello, 2003). Thirteen percent of the children came from a disadvantaged social background, distributed over the whole age range. This percentage corresponds with the percentage as known from the Dutch National Bureau of Statistics.

The normative sample existed of 324 children. We tested approximately the same number of boys and girls per age range. The ages ranged from three up to and including eleven years (see Table 1 for the age

distribution). Since the ToM Storybooks measure ToM aspects that should be developed at the age of five, we have tested more children before and right after this age, resulting in an under-representation of older children.

Table 1: *Age distribution of the normative sample being administered the ToM Storybooks (N=324)*

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

Material

The ToM Storybooks (see Blijd-Hoogewys et al., 2008; Serra et al., 2002) is a test that measures a variety of ToM components: emotion recognition, the difference between physical and mental entities (including tasks on close impostors and real-imaginary distinction), understanding that seeing leads to knowing, understanding of desires and beliefs (namely: standard belief, changed belief, not own belief, explicit false belief, false belief, inferred belief and inferred belief control).

There are 34 tasks, incorporated in short stories (for example tasks, see Appendix A). There are six storybooks in total. The stories are illustrated with full colour pictures and enlivened by the use of caressable patches of fur, toy doors that can be opened, and magnetized emotion faces that can be placed on the characters. The test takes 40 to 50 minutes, including a short break.

Scoring

There are 77 ToM test questions (the child has to predict an emotion or behaviour) and 18 ToM justification questions (the child has to explain the emotion or behaviour choice). The answers to the test questions are coded as correct or incorrect (1 or 0 points); the justification questions result in 2, 1 or 0 points, depending on the correctness of the mental state terms spontaneously used by the child. The evaluation of these justifications is founded on the category system used by Rieffe (1998), different categories

from Wellman (1990) and an exploratory analysis of the empirical data (see Appendix B).

The testing with the ToM Storybooks results in a maximum sumscore of 113 points (ToM sumscore), consisting of a maximum of 77 points for answers on the test questions (ToM-test score) and a maximum of 36 points for answers on the justification questions (ToM justification score).

Statistical method

Determining conversion curves

For both norming methods not only a model for the change in the average score was calculated but also a model for the change in variance. Since the average raw scores are expected to increase from some minimal value at the youngest ages to a plateau at the oldest ages, variability around the mean is likely to vary over the time axis.

The conversion curves were calculated with the help of the TableCurve 2D programme (Systat, 2000). In both cases, the best fitting curve (a monotonic or a Loess derived curve) as a function of age was chosen. First, the difference between the expected scores and the observed scores at all ages were calculated. The square of those differences produces the observed variance for every age in the intended age range. Next, given these observed variances, the expected variances at every age were estimated, also using a monotonic or a Loess derived curve. The square root of the expected variance at a particular age results in the expected standard deviation for that age. These scores form the basis of our norm score, the ToM-Q.

Checking for normality

Quotient-type norms require that the raw scores are symmetrically and preferably normally distributed, such that differences in terms of standard deviations clearly correspond with differences in frequencies (e.g. about two thirds of the subjects fall within one standard deviation of the middle). Since the scores increase with age according to the estimated function, the normality assumption must be checked for age-corrected scores, i.e. for the residuals. This was done by using the Shapiro-Wilk coefficient of normality and QQ-plots.

Calculating Z-scores and quotient scores

The calculation for the Z-scores and ToM Quotient scores was carried out using Visual Basic macros in Excel.

Raw ToM sumscores (S_o) were converted to Z-scores (Z_o), using the formula $Z_o = (S_o - S_m) / SD_m$, with S_o the raw score, S_m the model mean (the predicted mean score at age of S_o) and SD_m the model standard deviation.

The Z-scores were then converted to quotient scores, using the formula ToM-Q = (Z score*15) + 100 (Wechsler, 1981). These scores have an average of 100 and a standard deviation of 15. The minimum was set at 55, which is three standard deviations below the average.

RESULTS

Preceding analyses

Before calculating the standardized scores, we answered three preliminary questions: Do all items of the ToM Storybooks contribute significantly to estimating the ToM ability? Is the ToM sumscore useful for ordering subjects on their ToM ability? Are unisex norms sufficient or are gender specific norms required?

ToM sumscore as an estimation of ToM ability

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

The OPLM showed a good fit for the 95 ToM items (77 ToM test questions +18 ToM justification questions), except for three items, all concerning the inferred belief control task. For those items, a higher ToM ability did not result in a higher probability of giving a correct answer.

Therefore, those items were eliminated from the ToM test. The OPLM of the 92 remaining items revealed a good fit for all items.

The next question is how to combine the scores on the 92 items of a particular child to reflect his/her ToM ability. The OPLM enables us to estimate the ToM ability for each child. If the OPLM fits well, and only dichotomous items are involved, the (unweighted) sumscore orders subjects on their ability (Sijtsma & Hemker, 2000). Unfortunately, this property does not generally hold for polytomous items. However, the correlation between OPLM ToM ability estimate and the ToM sumscore appeared to be 0.997. Thus, the ToM sumscore and the OPLM ToM ability estimate yield approximately the same results for ordering the children on their ToM ability. Therefore, we may well confine the norming of the ToM Storybooks to the ToM sumscore (now with a maximum of 110), which is much easier for professionals in the applied field to compute and communicate.

Gender specific norms

On average, girls had slightly higher ToM sumscores on the ToM Storybooks than boys (see Table 2; M=71.71 versus M=68.73 respectively). The p-value of the difference, based on an independent samples t-test is .098; the variance hardly differed between both sexes (20.82 and 20.44) and is considered equal (Levene's test, p=.749).

When we divided the group in three age groups (n=87, <54 months; n=119, 54<78 months; n=118, ≥78 months), we found the gender difference to be significant for the youngest and oldest group (p=.05); and the variances within these two age groups were not equal (Levene's test, p=.01 and p=.05 respectively). Based on these results, we decided to generate separate norms for boys and girls. However, since the overall difference between boys and girls is relatively small (about 0.15 of the standard deviation), norms based on the total sample were also determined.

Table 2: *Gender differences on the ToM sumscore, measured with the ToM Storybooks, divided over three age groups*

Age	Boys		Girls	
	n	M (SD)	N	M (SD)
>54 months	**47**	**42.74 (12.28)**	**40**	**48.60 (18.04)**
36	14	36.00 (7.26)	13	35.08 (8.76)
42	18	38.94 (6.77)	16	47.00 (13.88)
48	15	53.60 (14.16)	11	66.91 (16.64)
54-78 months	**60**	**73.00 (11.97)**	**59**	**71.25 (13.47)**
54	16	66.13 (11.70)	13	63.08 (13.39)
60	14	73.50 (12.34)	16	64.56 (8.51)
66	17	75.76 (7.47)	16	78.56 (13.51)
72	13	77.31 (14.13)	14	78.14 (10.57)
≥78 months	**60**	**84.82 (11.81)**	**58**	**88.12 (9.33)**
78	18	72.83 (9.62)	12	79.25 (9.61)
84	15	86.40 (8.16)	16	88.69 (8.59)
96	14	90.07 (8.09)	12	91.42 (7.74)
120	13	93.92 (8.12)	18	91.33 (7.32)
all ages	**167**	**68.73 (20.82)**	**157**	**71.71 (20.44)**

Note. n= number of subject; M= average ToM sumscore; SD=standard deviation.

Norming approach 1: the monotonically rising regression

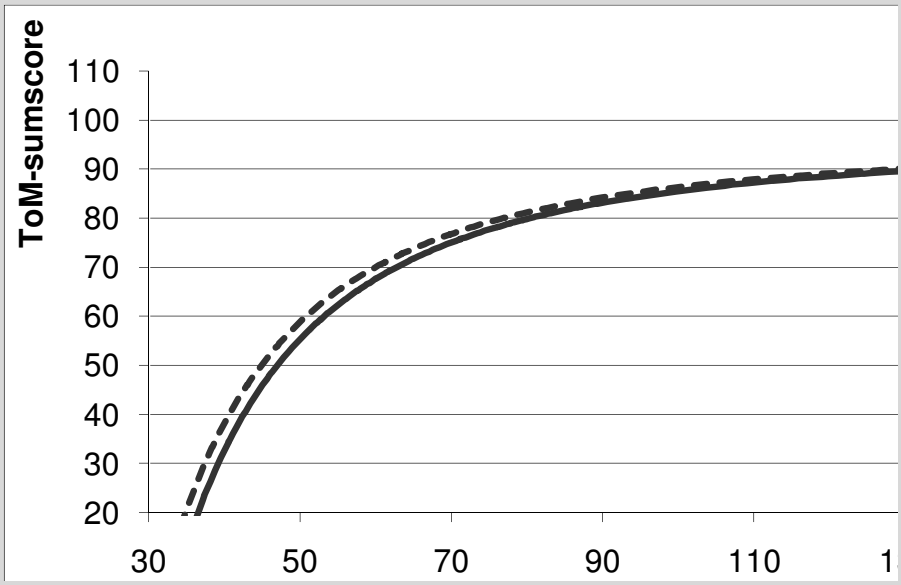
With regard to the conversion curves, the best fitting (highest r^2 adjusted) and simplest monotonically rising curves as a function of age were chosen (see Figure 1). An additional assumption was made, namely that the ToM sumscore must become stable at the age of 12 and cannot increase any further. We also controlled for the estimated score at the ultimate age not to be higher than the maximum possible sumscore (110; the maximum score that the chosen model yields is 95. rounded off to the nearest integer).

For the boys, the estimated ToM-scores were calculated with the formula: $y = 95.55 - 100647.97/age^2$. The r^2 adjusted was .71. The standard deviations were calculated with the formula: $y = \sqrt{(251.63-15.37*\sqrt{age})}$.

For the girls, the estimated ToM-scores were calculated with the formula: $y = 95.50-91946.74/age^2$. The r^2 adjusted was .67. The standard deviations were calculated with the formula: $y = \sqrt{349.45}*e^{(-age/74.06)}$.

Before calculating the Z-scores and Q-scores we checked the normality assumption for the score distribution. The p-value of the Shapiro-Wilk coefficient of normality of the residuals equals 0.04, which means that the distribution of the residuals may be rejected as a sample of normal deviates with a significance level of 0.04. However, visual inspection of the residuals using a QQ-plot suggested that the distribution was sufficiently close to a normal distribution. The maximum of the distribution was moderately shifted to the right. However, we consider the deviation from the distribution not serious enough to exclude norms in the form of Z-scores and Q-scores. The norm scores were calculated for boys and girls separately.

Figure 1. A monotonic regression fit of ToM sumscore data plotted versus age, for boys (black line) and girls (dashed line). These two regressions are not statistically different, indicating that the simple fit does not distinguish any difference between boys and girls as they age.



Norming approach 2: the Loess curve

Norms for the ToM sumscores were obtained by using Fourier-series tenth-order polynomials that provide the best fitting equations for curves that are based upon a Loess 30% window model (see Figure 2).

The Fourier series polynomials were used to calculate the expected score for every age in the age range. The r^2 was .77 for the boys and .71 for the girls. Given these observed variances, we used the same Loess 30% window technique to estimate the expected variance at every age. For, it is assumed that not only the average scores but also the variance may change non-linearly, eventually showing developmental regressions and plateaus (see Van Geert & Van Dijk. 2002). The Loess curves were transformed into Fourier-series polynomials.

Figure 2. A Loess regression fit of ToM sumscore data plotted versus age. for boys (black line) and girls (dashed line). These two regressions are statistically different; the Loess distinguishes differences between boys and girls as they age.

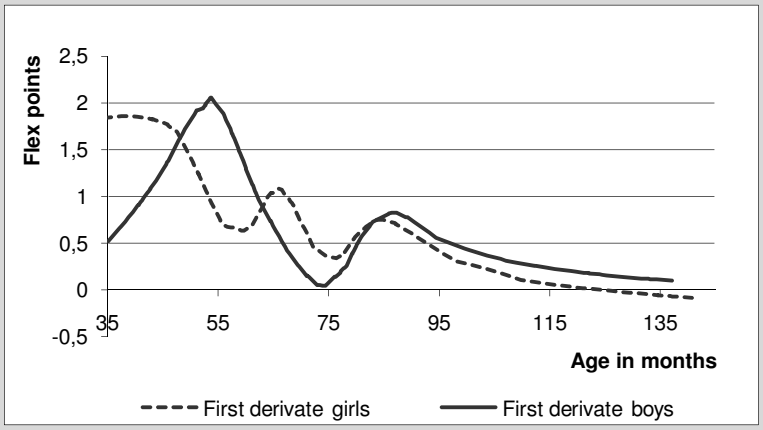
In order to check whether the Loess-fitted models can provide a feasible starting point for the calculation of quotient scores, the Shapiro-Wilk coefficient of normality of the residuals was calculated. Its p-value is equal to

0.37, which implies that the assumption that the residuals are drawn from a normal distribution should not be rejected. Nevertheless, the peak of the distribution is also slightly shifted to the right, similar to the residuals of the monotonically rising regression model. The QQ-plot showed that the residuals were normally distributed. Norm scores were calculated for boys and girls separately.

Gender differences

Only the Loess curves were able to capture a gender difference (compare Figures 1 and 2). The differences between both genders were mainly obvious in the youngest age group (Figure 2). If we look at the first derivate as an indicator of change, we see that the peaks and valleys are quite different for boys and girls (Figure 3). Girls show three distinct moments of change/acceleration; boys show only two periods of change, with their first change being later. After the age of approximately 78 months (6.5 years), the curves for boys and girls merge.

Figure 3. A Loess curve of the first derivate, as an indicator of change, of the ToM Quotient scores plotted versus age, for boys (black line) and girls (dashed line). The girls show three distinct peaks, while boys only show two. There are distinct gender differences up to the age of 78 months (6.5 years).



A comparison between the two standardization approaches

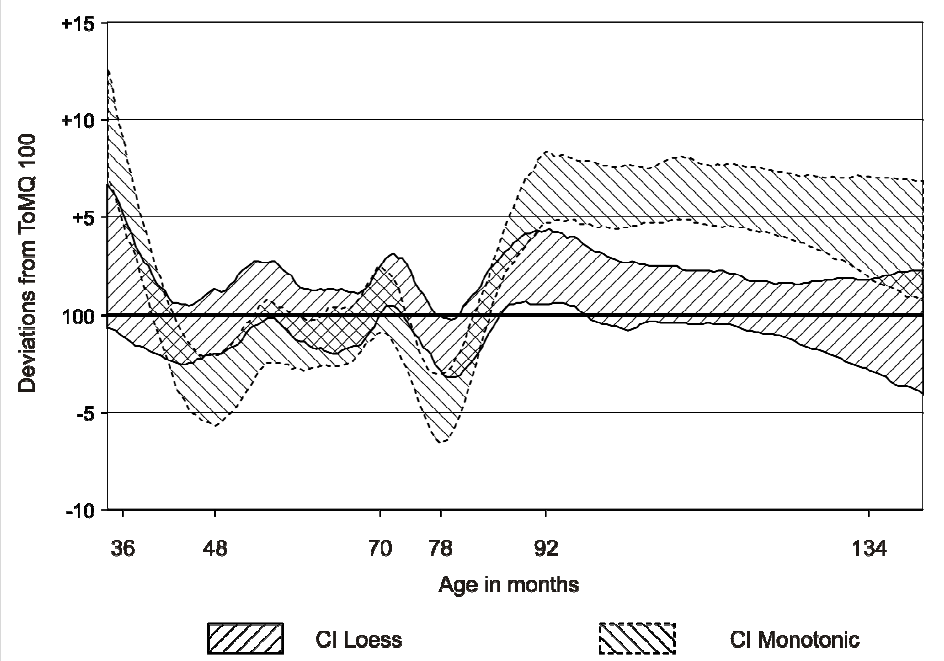
Quotient scores derive their applied functionality from the fact that they have a constant average, a constant standard deviation and a constant frequency distribution irrespective of age.

In order to compare the two standardization approaches, we first checked the distribution of ToM-Q scores, which should be normal. Visual inspection of QQ-plots showed that both distributions were sufficiently close to normality. However, the Shapiro-Wilk coefficient of normality had a p-value smaller than 0.001 in the case of the monotonous curve method and equal to 0.11 in the case of the Loess method. Thus, the normality assumption should be rejected in the first case and not – or less clearly so – in the second.

A more relevant comparison between the two models can be achieved by calculating the average deviation from the average of 100 and the standard deviation of 15 from the quotient values of the standardization sample. The deviations for the expected norm scores were 3.03 and 0.95 for the monotonous and Loess standardization models respectively and 0.95 and 0.35 for the expected standard deviations. The differences between both methods were statistically significant (Monte Carlo analysis. $p< 0.001$). The standardization method based on the Loess procedure was superior.

A final way to compare the two standardization approaches is by visually comparing the averages of all quotient scores over time. This was done by fitting the quotient scores resulting from both methods by means of a Loess 30% window model with age as predictor variable. The expectation was that the curves showing the average of the scores over age should be arbitrarily close to 100. In order to allow a more reliable comparison between the curves resulting from both approaches 60% confidence intervals were calculated by means of a bootstrapping procedure. Inspection of the curve (see Figure 4) taught us that the Loess model approaches the average more consistently than the monotonic model and moreover that the confidence intervals of the curves show only minor overlap.

Figure 4. Comparing the confidence intervals of the deviations from the average ToM quotient score of 100 of two norming methods, the monotonic regression model (lines slanting to left) and the Loess norming model (lines slanting to right) plotted versus age.



Note. CI= Confidence Interval.

Using the ToM Storybook in the applied setting

A quick-and-dirty method

Thus far, the norming procedures were based on both the 74 binary ToM test questions and the 18 ordinal ToM justification questions. In practice, a quick-and-dirty method can be obtained using only the 74 binary ToM test questions (for. justifications take a lot of time to get evaluated). However, we advise practitioners to use the data including the justifications, since these can give additional insight in the ToM problems of children. Norms based on solely the binary (quantitative) data can be requested from the authors.

Age equivalents

We wrote an Excel Visual Basic macro to calculate the ToM age-equivalent of a child, i.e. the age for which the model mean equals the child's observed score (equal to a ToM of 100). For instance, a boy with a ToM sumscore of 64 at the age of 66 months has a ToM-Q of 89. This boy functions at the level of a 57 month-old child; he displays a ToM-ability delay of nine months.

Confidence intervals

To calculate confidence intervals for the ToM Storybooks, we first fitted a Loess (30% window) regression model for each separate item with the total test score as 'predictor'. The Loess model of each separate item specifies the probability curve of the item, i.e. each possible total test score corresponds with a specific probability that this item will be answered correctly (this procedure is similar to specifying a kind of IRT model). Second, for each possible sumscore on the test (i.e. all possible *presumed* sumscores), we calculated the corresponding distribution of observable sumscores based on the Loess probability models of each separate test item.

In order to calculate the confidence interval of an actually *observed* score, we made a weighted sum of all those probability distributions. the weights being determined by the likelihood that the observed score occurs in each of the probability distributions. Thus, the probability distribution D_o of presumed scores that range from a score s_{min} to a score s_{max} is the sum of the distributions corresponding with each possible, presumed score s. D_s. multiplied by the likelihood of the observed score o given a presumed score s. $p_{o \mid s}$:

$$D_o = \sum_{s_{min}}^{s_{max}} D_s \cdot p_{o|s}$$

An attractive property of the distributions and related intervals calculated with this method is that they are not symmetric where symmetry is not readily expected, namely in the region of very high and very low scores. For instance, the extreme low score of 20 has a 50% CI between 21 and 26 and a 90% CI between 17 and 31, thus expressing the fact that a score of 20 is probably a matter of low ToM knowledge in combination with bad luck (for details about the procedure and the underlying simulations of the score distributions. we refer to www.vangeert.nl).

CONCLUSION AND DISCUSSION

In this article, we demonstrated the statistical strengths of a comprehensive ToM test and argue for its adoption in the applied and clinical setting. The newly developed ToM Storybooks were administered and norms were calculated in two distinct ways. The first norming method was based on the idea that a cognitive ability, such as ToM, must increase monotonically in the population of typically developing children, assuming that the underlying ability level cannot regress or decline. The second norming method was based on a strongly developmental and performance-oriented approach, which accepts the occurrence of plateaus and even temporary developmental regressions in the performance level of an ability.

Preference was given to norm scores in the form of quotients, primarily because these are customary in applied contexts and easy to use and interpret. Differences in ToM sumscores were found between boys and girls, which are in accordance with findings on false belief tasks (e.g. Charman. et al.. 2002; Cutting & Dunn. 1999). Because of these gender differences, gender-specific norms were calculated. If we had taken boys and girls together as one group, this would mainly have lead to an underestimation of the ToM ability of young boys. A similar difference in gender, where boys are slower than girls, has also been reported in other developmental areas (e.g. Luotonen, 1995).

The comparison between boys and girls turned out to reveal developmental differences in the curves in the case of the Loess method but not in the monotonic method. Further comparison of the two norming methods revealed the Loess fitting method to be superior. First, the requirement that quotient scores need normally distributed scores could be met by first correcting the scores for developmental trend (either on the basis of the monotonically or Loess methods). In doing so, the Loess method provided a considerably better approximation to the normal distribution than the monotonically rising fit. Second, the expected average of 100 was at each age significantly better reached by the Loess method, also supported by the visual comparison of the confidence intervals of the deviations from 100. Thus, it can be concluded that the quotient scores based on the Loess method, which more faithfully follows plateaus and temporary developmental regressions in the raw scores, are superior over the scores based on the assumption of monotonic increase.

One of the restrictions of this research is that fewer children in the older age regions were tested, which implies a reduction in reliability. This

is due to the fact that the test is primarily intended for younger children, up to five or six 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. In future, additional tasks should be included, like for instance second-order belief tasks, which are better suited for older children (see for instance Hughes et al., 2000).

Next to that, we presented age equivalents and confidence intervals. They were calculated with a new method based on simulations of observed score distributions, given certain presumed scores. Confidence intervals were determined by calculating presumed score distributions for given observed scores. It should be noted that the distributions on which the confidence intervals were based should be treated in ways similar to Bayesian marginal (or prior) probabilities, i.e. as probabilities prior to any other knowledge about a child's test score than the test score itself (and the implicit assumption that testing has occurred under normal conditions). If more information becomes available, e.g. a retest score, the intervals should be updated in ways consistent with that new information. Note, however. that retesting is likely to result in an increase of the score, due to the learning effects that appear to be quite substantial, particularly in young children. This was demonstrated by Blijd-Hoogewys and colleagues (2008) using the ToM Storybooks in both typically developing children and children with autism spectrum disorders.

In summary, a norming procedure based on a Loess smoothing method has resulted in norms for a broad age range that enable the practitioner to calculate quotient scores with a consistent, observed average of 100 and a standard deviation of 15, in spite of developmental phenomena such as temporal developmental regressions in the raw scores. In addition, age equivalents and reliable confidence intervals have been provided, in order to further enhance the measurement of ToM in both typically developing children and clinical populations. The two methods proposed - a non-linear way of calculating norm scores and the utilization of bootstrapping procedures in calculating confidence intervals - can be considered innovative and effective contributions to future norming studies.