Factorial structure of the 'ToM Storybooks': A test evaluating multiple components of Theory of Mind

Bulgarelli, D., Testa, S., & Molina, P. (2015). Factorial structure of the 'Tom storybooks': A test evaluating multiple components of Theory of Mind. *British Journal of Developmental psychology*, 33(2), 187-202 [DOI: 10.1111/bjdp.12062]

Abstract

This study examined the factorial structure of the ToM Storybooks, a comprehensive 93-item instrument tapping the five components in Wellman's model of ToM (emotion recognition, understanding of desire and beliefs, ability to distinguish between physical and mental entities, and awareness of the link between perception and knowledge). A sample of 681 three- to eight-year-old Italian children was divided into three age groups to assess whether factorial structure varied across different age ranges. Partial credit model analysis was applied to the data, leading to the empirical identification of twenty-three composite variables aggregating the ToM Storybooks items. Confirmatory factor analysis was then conducted on the composite variables, providing support for the theoretical model. There were partial differences in the specific composite variables making up the dimensions for each of the three age groups. A single test evaluating distinct dimensions of ToM is a valuable resource for clinical practice which may be used to define differential profiles for specific populations.

Keywords: Theory of Mind, ToM Storybooks, confirmatory factor analysis, comprehensive test

Factorial structure of the "ToM Storybooks": a test evaluating multiple components of Theory of Mind

Background

Theory of Mind (ToM) is the social cognitive ability to attribute mental states to oneself and others, and to use these attributions to understand, predict and explain one's own behaviour and that of other people (Mitchell, 1997). Research on ToM development has primarily investigated children's ability to carry out focal tasks, the most widespread being the false belief task (Wimmer & Perner, 1983). Typically, 3-year-old children systematically fail the task, whereas 4-year-olds provide a correct response. Several versions of the task have been developed (for instance: Baron-Cohen, Leslie, & Frith, 1985), and a meta-analysis of studies conducted using false belief tasks has found a consistent trend in children's responses (going from below chance to above chance probability of providing a correct answer in the course of the preschool years), independently of the type of task, type of questions, nature of the protagonist or the object, etc. (Wellman, Cross, & Watson, 2001; Wellman, 2012). Moreover, the responses of children from the United States, United Kingdom, Korea, Australia, Canada, Austria and Japan varied similarly as a function of age, suggesting a basic universal developmental pattern for ToM (Wellman et al., 2001; Wellman, 2012). However, within an overall universal pattern, false belief processing also displays developmental specificities as a function of cultural and linguistic features: for example, Junin Quechua children from Peru (Vinden, 1996), Samoan children (Callaghan, et al., 2005; Mayer & Trauble, 2012), Mofu children from Cameroon and Tolai children from Papua New Guinea (Vinden, 1999) have been found to show delays in passing false belief tasks compared to Western children. Differences have also been found among Western European countries, with Italian children showing a slight delay relative to English children (Lecce & Hughes, 2010).

Nevertheless, ToM is more than false belief understanding. Tager-Flusberg & Sullivan (2000) distinguished between social-perceptual and social-cognitive components of children's

understanding of the mind, specifically linked to distinct neurobiological areas. The first component concerns the ability to distinguish between persons and objects and to infer people's mental states from their facial and body expressions; this competence emerges very early in development and appears to be specifically impaired in Williams syndrome. The social-cognitive component regards understanding the mind as a representational system: this is the conceptual dimension of ToM, which emerges during the preschool years; it is usually evaluated via false belief tasks, on which children and adults with autistic syndrome perform poorly (Baron-Cohen, 2000).

More recently, this conceptual dimension of ToM has also been defined as explicit or declarative ToM, to distinguish it from a more non-verbal and implicit (procedural) competence, observed in early infancy via a number of alternative paradigms (Low & Perner, 2012): anticipatory look (Clements & Perner, 1994), violation-of-expectation (Onishi & Baillargeon, 2005; Surian & Geraci, 2012; Yott & Poulin-Dubois, 2012), and interpretation in the course of interaction of other people's intentions on the basis of their prior knowledge (Buttelmann, Carpenter, & Tomasello, 2009). The relations between implicit and explicit competences still represent a point of debate in the literature (for a review see Caron, 2009; Low & Perner, 2012; Sodian, 2011). Moreover, longitudinal research is very scarce, and the existing results are puzzling: for instance, Thoermer, Sodian, Vuori, Perst, and Kristen (2012) found a high level of correlation between an implicit (18 months) and explicit (48 months) location-false-belief task, but no correlation between the same implicit task and a content-false-belief task at 48 months.

Within the explicit ToM framework, according to Wellman (1990), basic ToM comprises five different types of ability: recognizing emotions, making a distinction between physical and mental entities, appreciating that perception leads to knowledge, understanding how desires affect behaviour and understanding how beliefs affect behaviour. During the preschool and school years, children refine and develop their knowledge about each of these ToM components. On the basis of empirical research data, Flavell, Miller and Miller (1993) argued that children's understanding of the mind improves across the acquisition of five concepts, and that basic knowledge of each step is

required to reach the next one. At a first level, the child realizes that the mind exists, and starts using mental and emotional lexicon; then, the child learns that the mind is connected to the physical world, which may be perceived. At a third level, the child realizes that the mind is private, separate and different from the physical environment and, at a later stage, it understands that the mind may or may not be accurate in representing objects and events, a realization that is critical to success on false belief tasks. Finally, children of school age mature a more sophisticated understanding of the mind as actively mediating both people's interpretation of reality and the mental states they experience.

Therefore ToM ability emerges in the early years and improves during the preschool and primary school stages. Given that the individual components of ToM may display different developmental trends, comprehensive instruments allowing the different ToM competences to be evaluated simultaneously - such as the ToM Tasks (Wellman, Fang & Peterson, 2011; Wellman & Liu, 2004), the ToM-Test (Muris et al., 1999; Steerneman & Meesters, 2009) and ToM Storybooks (Blijd-Hoogewys, van Geert, Serra, & Minderaa, 2008) - are of great value. The ToM Tasks by Wellman and Liu (2004) assess seven ToM components selected via a meta-analysis of the most widespread tasks used to evaluate this ToM ability in preschool children; a Rasch analysis was performed on data collected with pre-schoolers showing that the tasks were hierarchically ordered from the easier ones (diverse desire, diverse belief, knowledge access) to the more difficult ones (content false belief, explicit false belief, belief emotion, real-apparent emotion). The ToM-Test (Muris et al., 1999) is designed for children from 5 to 12 years and assesses precursors of ToM (perception and imitation, emotion recognition, pretence and physical-reality distinction), elementary ToM (belief and false belief reasoning) and advanced ToM (second order belief understanding, understanding of complex humour); children's performance on the test improves with age, with the advanced ToM tasks remaining more difficult at all ages.

The ToM Storybooks is a test that taps into all five components of the ToM model proposed by Wellmann, and was originally designed and standardized in the Netherlands (Blijd-Hoogewys et

al., 2008). The test was subsequently translated into Italian and data collection with Italian samples began in 2005, on the basis of an international agreement between the Dutch authors of the test and Italian researchers (first and last authors). At the time, no such comprehensive test for the evalutation of ToM was in use in Italy. The ToM Storybooks was chosen for several reasons. First, the test had been devised following the dynamic system approach (Thelen & Smith, 1994; van Geert, 2003): differently from the ToM Tasks and ToM Test, in the ToM Storybooks four of the five components are evaluated via two or more similar tasks set in different story contexts, thus enabling more precise and stable measurement (Hughes et al., 2000). Second, the test comprised both yes/no questions and justifications, so as to evaluate whether children spontaneously refer to mental states in order to explain the behaviour of others. Third, the test was designed to cover a broad age range, from 3 to 8 years and, fourth, the Dutch version had already displayed good reliability and validity (for details, see Blijd-Hoogewys et al., 2008).

Preliminary data for the Italian standardization of the ToM Storybooks were collected from a sample of 466 typically developing children (223 females; age: 3-8 years), between 2005 and 2010. The ability to provide correct responses was found to increase regularly with age. The tool proved to be reliable: Cronbach's alpha for the dichotomous items was .80 and Cohen's K for the justifications was .89, similar to the values obtained in the Dutch test. No differences emerged as a function of gender, presence of siblings in the family, having a foreign parent or speaking languages other than Italian in the home (Molina and Bulgarelli, 2012). Furthermore, the test showed good validity in discriminating children with Pervasive Developmental Disorder (PDD) in the Netherlands and this result was replicated in a preliminary analysis of data from a small sample of Italian children with PDD (Bulgarelli, Molina, & Salomone, 2011).

Today, the ToM Storybooks has been translated into Italian, French (Thommen, Cartier-Nelles, Guidoux, & Wiesendanger, 2011) and Finnish (Vesterinen, 2008), and it has been standardized on two European populations, Dutch children (Blijd-Hoogewys et al., 2008) and Italian children (Molina & Bulgarelli, 2012). This is an interesting feature of test, for two main

reasons. First, the availability of normative data represents an advantage for both research and clinical use of the tool and, to our knowledge, no other comprehensive tests such as the ToM Tasks and the ToM Test have been standardized to date. Moreover, the ToM Storybooks has been standardized on two European populations, Dutch and Italian children, belonging to two different cultural European clusters, the Germanic and the Latin groups, respectively. Given that cultural belonging to these two clusters has been linked to differences in the development of both ToM and emotion understanding (Bulgarelli, Blijd-Hoogewys, van Geert, & Molina, 2011; Lecce & Hughes, 2010; for a discussion on the European clusters see Molina, Bulgarelli, Henning, & Aschersleben, 2014), this double standardization is a strong point of the ToM Storybooks: the Dutch norms could be more representative for the Northern European countries, while the Italian norms could be more representative for the European countries in the Mediterranean area.

To explore the factorial structure of the instrument, Blijd-Hoogewys and colleagues (2008) carried out a Simultaneous Component Analysis (SCA-P) on 21 aggregations of theoretically similar items termed composite variables (briefly described in the Appendix). They obtained a five-factor solution accounting for 53.8% of variance that replicated Wellman's model reasonably well, although, as noted by the authors, two composite variables did not fit the factorial structure (desire action, DA and seeing is knowing, SKW, see Appendix for the full list of abbreviations), while two other composite variables (Mental Physical senses and Close Impostors future) displayed high loadings on two or three components.

Aims

The overall purpose of the current study was to further investigate the factorial structure of the ToM Storybooks. A more precise understanding of which dimensions of Wellman's model the instrument effectively assesses would represent a preliminary step towards developing subscales that, along with the total score, could yield a richer assessment of children's theory of mind competences. As mentioned in the Introduction, the five-component- solution obtained in the Dutch study was only partly satisfactory given that two composite variables did not fit the model while

two others had high loadings on two or three components. Furthermore, in the Dutch analysis the emotion component of Wellman's model was split into two dimensions. As noted in the literature (e.g. Bandalos & Finney, 2001) if item parcels (i.e., the composite variables in our case) are lacking in unidimensionality, this can obscure the true factorial structure of the data. The aims of this study were to empirically determine unidimensional composite variables and assess their fit with Wellman's model by means of confirmatory statistical analysis.

Method

Participants

The sample comprised 681 three- to eight-year-old children (*M* = 68 months, *SD* = 19 months, *range* = 31 – 103 months, 353 females), from a large Northern Italian city and its suburbs. Of these, 443 children had been involved in an earlier study by Molina and Bulgarelli (2012), for which data were collected between 2005 and 2010. In 2011 and 2012, a further 238 children were tested, in very similar conditions. Children were recruited at infant-toddler centres, preschools and primary schools, and in a few cases via personal contacts; individual administration of the test required about 30-40 minutes depending on the child's age. All families provided written informed consent for their child's participation in the research. Five children (4 boys and 1 girl, average age = 39 months) were excluded from the analysis, because they did not fully complete the test.

Parents and teachers reported that all children spoke Italian and were not affected by medical, psychological or developmental disorders. Italy was the birthplace of 89.9% of the children (N = 612), while 2.9% had been born in other countries (N = 20); this information was missing for 7.2% of the sample (N = 49).

With regard to family composition, 59.8% of children had at least one sibling (N = 407), 31.1% were only children (N = 212), while this information was missing for 9.1% of the sample (N = 62); the percentage of only children in the general Italian population was higher (41.1% for mothers aged 25-54; Istat, 2006) than in the sample. In terms of education, 32.5% of mothers had completed middle-school (N = 221), 36.6% held a high-school diploma (N = 249) and 19.4% were

graduates (N =132); again, 32.9% of fathers had completed middle-school (N = 224), 35.1% held a high-school diploma (N = 239) and 16.7% were graduates (N =114). The percentage of parents' with middle school diplomas was lower than in the general Italian population (59.7% for women and 57.8% for men; Istat, 2003), whereas the percentage of university degrees was higher (9.3% for women aged 25-65 and 10.2% for men aged 25-65, Istat, 2003). Level of education was not reported for 79 mothers (11.6%) and 104 fathers (15.3%). Regression models were performed and the ToM scores of the children for whom socio-demographic data was missing did not systematically differ from the ToM scores of children for whom complete demographic profiles had been collected.

To test the factorial structure of the Tom Storybooks instrument across different age ranges, the sample was divided into three age groups: 206 three- and four-year-old children (*M* = 46 months, *SD* = 5, *range* = 31 – 54, 113 females); 243 five- and six-year-old children (*M* = 67 months, *SD* =7, *range* = 54 – 78, 120 females); 232 seven- and eight-year-old children (*M* = 90 months, *SD* = 8, *range* = 78 – 103, 120 females).

Instrument

The ToM Storybooks is made up of six full-picture books telling stories about a boy called Sam. Each book recounts an adventure of Sam's (Sam going to the swimming pool, visiting his grandparents, etc.) and contains 5 or 6 tasks assessing one or more ToM components. Altogether the test comprises 34 tasks. Each task gives rise to 3 or 4 questions yielding a total of 93 items, 75 of which are scored dichotomously (correct or incorrect) and 18, termed justification questions and designed to verify whether the child spontaneously refers to mental states, are scored on a three-point scale (completely correct, partially correct, incorrect). Each task therefore comprises both quantitative and qualitative questions, yielding a "quantitative" score varying from 0 to 75 and a "qualitative" score varying from 0 to 36. Total ToM score is obtained by summing the quantitative and qualitative scores (max. score: 111). In the present study, Cronbach's alpha for the dichotomous

items was .92 for the whole sample, and .77, .85 and .80 in the younger, middle and older age groups, respectively.

Statistical analysis

The analysis was performed in two steps. In the first step, 75 items were analysed using Partial Credit Models (PCMs, Write & Masters, 1982) on the data from the entire sample (n=681) in order to empirically obtain unidimensional scales. PCM, and Rasch models in general, are particularly useful for analysing scales in which some items could present ceiling or floor effects, as in our study. A Rasch model for polytomous items was adopted because the score for each of the 18 justification items (0-2) was added to the score of the associated dichotomous item, yielding items with scale categories ranging from 0-3. Dimensionality was assessed by performing principal component analysis on the residual matrix of the PCM. As a rule of thumb, scales for which the first eigenvalue was less than or equal to 2 were considered unidimensional (Linacre, 2009). Item functioning was assessed by checking that point-biserial correlation was positive and that Infit and Outfit statistics were within acceptable ranges, i.e. .5-1.5 (Linacre, 2009; Wright & Masters, 1982). Differential item functioning (DIF) for gender and age groups was also evaluated; given the high number of significance tests involved, a conservative significance level of .001 was adopted in place of the conventional .05 level applied to the other statistical analyses conducted as part of in the present study. For the 18 polytomous items, category structure was analyzed and when disordered thresholds occurred, items were rescored by collapsing categories. All the first-step analyses were carried out using Winsteps 3.68.2.

In the second step, CFA models were performed separately for the three age groups on the variance-covariance matrix of the composite variables obtained in Step One (PCM scales). Given that the data violated the multinormality condition as evaluated by Mardia's multivariate test, the Robust Maximum Likelihood method (MLR, Muthén & Muthén, 2004) was adopted. The following criteria were used to evaluate the model's goodness of fit: RMSEA < .08; CFI > .95; SRMR < .08 (Hu & Bentler, 1995; 1999). Mplus 6.11 was used for these analyses.

For each CFA dimension a summed score was then calculated and one-way ANOVAs were performed to test for linear trend in mean values as a function of age.

Results

Step One: empirical identification of the ToM composite variables

Starting from the results of the SCA-P analysis of the Dutch study and given the need for aggregations with a reasonable number of items (8-10 items minimum for each model tested), the following seven groups of items were hypothesized and analysed using PCMs (in brackets, see the correspondence with the Dutch composite variables, which are described in detail in the Appendix): Emotion (emotion naming + emotion recognizing), Belief-Emotion (recognition + naming), Desire (action + emotion recognition + emotion naming), Mental-Physical (senses + others + future), Close Impostor (senses + others + future), Real-Imaginary, Belief-Action (standard belief + changed belief + false belief + not own belief + non-belief + inferred belief control).

None of the seven groups of composite variables satisfied the eigenvalue criteria and they were split into new scales until the criteria were met. The algebraic sign of the loadings on the first component was used to group the items into the new scales. The resulting 23 PCM scales (listed in Table 2) displayed good item fit statistics: point-biserial correlations with the total score were all positive, Infit and Oufit statistics were within acceptable ranges and DIF analysis showed no gender bias. Seven out of the 18 polytomous items presented disordered thresholds and were rescored, while 13 items were not invariant across age groups (PCM results are not shown here).

Step Two: CFA models

Average scores on the 23 PCM scales were calculated and the resulting variables rescaled to a 0-1 range. The polytomous item referred to as Seeing is Knowing was also rescaled to 0-1 and CFA was carried out on the 24 variables, i.e. the 23 empirical composite variables (ECVs) and the single item Seeing is Knowing.

For each of the three age groups, Wellman's five-dimension model was specified in the following manner: Emotion Recognition (Emotion) measured by ECVs 1 through 7 of Table 2;

Desire Understanding (Desire) measured by ECVs 8 through 13; Distinction between Physical and Mental Entities (Mental-Physical) measured by ECVs 14 through 21; Belief Understanding (Belief) measured by ECVs 22 and 23. The single polytomous item Seeing is Knowing was used as sole indicator of the Perception Leading to Knowing dimension (Perception Knowledge). For the second and third age groups, ECVs displaying a ceiling effect (marked with asterisks in Table 2) were not included in the model.

The theoretical model did not display good fit for any of the age groups, leading to convergence problems in relation to age group 1. A refined model – in which the ECV related to happiness in the Desire factor were considered to be an indicator of the Emotion factor and the ECV Desire Action was considered to be an indicator of Belief factor – was taken as the initial estimable model for all age groups. The fit measures for this model are summarized in Table 1. RMSEA and SRMR fit statistics were good but the CFI was far from the acceptance threshold, especially in age groups 3 and 1. The initial model was therefore respecified by eliminating the ECVs with loadings that did not reach the .05 level of statistical significance or with loadings of the wrong algebraic sign. For age group 3, the Perception Knowledge single-item factor was eliminated because it was not correlated with the other dimensions. The final model obtained had acceptable fit measures (Table 1): in age groups 2 and 3, CFI values were under .95 but in any case over .90.

Table 1 about here

As shown in Table 2, for the Emotion factor the ECVs related to fear, sadness, happiness and anger were good indicators of ToM in younger children, while ECVs related to surprise were good indicators for older children. The Desire factor was made up of sadness tasks only, with the happiness tasks loading on the Emotion factor and the desire-generating behaviours (DA) on the Belief factor. Composition of the Mental-Physical factor varied greatly as a function of age: only the ECV “MP tasks 1 & 3” was common to all three age groups; moreover, the Close Impostors ECVs showed satisfactory loadings as a function of the specific sense involved in perception and the child’s age: smelling was a good indicator for age groups 2 and 3, touch for age groups 1 and 2,

whereas seeing was a weak indicator and only for age group 2. The three ECVs related to the distinction between real and imaginary phenomena were the poorest indicators of this factor, given that most of these ECVs were non-statistically significant, or displayed a ceiling effect. Concerning the Belief factor, the standard belief and false belief ECV was a good indicator for all age groups, while the ECV composed of the other belief tasks was mainly a good measure for age group 3. Finally, the single-item factor Perception Knowledge was not taken into account for age group 3 because it was not correlated with the other ToM factors. Emotion, Mental-Physical and Belief factors were highly correlated with one another for all age groups, while Desire was the least correlated dimension (Table 3).

Table 2 and Table 3 about here

Means and standard deviations of the CFA dimensions

Table 4 reports the means of the CFA dimensions for six-month interval age classes. As assessed by the one-way Anova, the age-linear trend was always statistically significant for the Belief and Mental-Physical dimensions, while the Desire dimension varied significantly by semester only between 3.0 and 4.5 years, the Emotion dimension only for the semesters between 3 and 6.0 years and the Perception Knowledge dimension only between 4.5 and 6.0 years.

Table 4 about here

Discussion and conclusion

The aim of the current study was to further investigate the factorial structure of the ToM Storybooks, a comprehensive test tapping five ToM components, with a sample of Italian children. Given that the test comprises up to four different versions of the same task, our analysis was first designed to empirically identify how the test items grouped into composite variables. The empirical identification of unidimensional composite variables via PCM analysis showed that the principal item organizers were type of emotion (anger, fear, happiness or surprise), specific nature of the task (in terms of the level of difficulty of the syntactic structure of the sentence and the related cognitive effort – for example the use of positive vs. negative sentences – as in the case of the desire-emotion

tasks), type of action involved in perception (smelling, seeing or touching), and standard belief and false belief tasks vs. other belief tasks.

The CFA on the empirical composite variables provided support for the five dimensions that we had hypothesized on the basis of both the factorial structure found for the Tom Storybooks in the Dutch study, and Wellman's model. Four out of five dimensions were common to the three age groups: Emotion, Desire, Mental-Physical and Belief. However, there were partial differences in the set of composite variables making up these dimensions for each of the age groups: in fact, some items displayed a ceiling effect, as expected given that the test was developed to cover a wide age range (3-8 years); other items were not good indicators given that they were found not to be invariant across age groups in the PCM analysis.

The key dimensions measuring ToM were Mental-Physical, Belief (including False Belief) and, up to 6 years of age, Emotion. In fact, these dimensions were highly correlated with one another, and children's scores on them increased with age, as theoretically expected. The Desire dimension showed lower and sometimes non-significant correlations with other dimensions: from four-five years onwards knowledge about the role of desire appeared to have been consolidated, with scores no longer increasing with age. This result is coherent with the theoretical model, in which understanding of desire is expected to be present from three years onwards (Wellman, 1990). The Perception Knowledge dimension displayed a similar trend, a plausible result given that this item assesses a basic competence usually acquired at a relatively early age.

Some interesting differences emerged with respect to the Dutch validation study (Blijd-Hoogewys et al., 2008). The theoretical criteria informing the distinctions between "recognition" and "naming" emotions and among "senses", "others" and "future" in the Mental-Physical and Close Impostor tasks were not supported by the PCM analysis. The empirical construction of the composite variables produced more theoretically plausible dimensions. For example, in the Dutch study, the factor covering emotions was split into two mildly correlated latent dimensions, although both assessed competences related to the understanding of emotional state. Moreover, the desire

action composite variable loaded on the Mental-Physical factor, whereas in this study it loaded on the Belief factor: this last solution appears to be more appropriate, because both the desire and belief tasks deal with a mental state generating behaviours. Finally, we did not find confirmation of the invariance across age groups identified by Blijd-Hoogewys and colleagues when they imposed the same pattern of loadings across the three age groups. Although four out of the five components in Wellman's model, namely emotion, desire, mental-physical distinction and beliefs, could be differentially assessed at any age ranging from 3 to 8 years old, desire and emotion scores did not discriminate among the oldest children. Furthermore, only some of the composite variables in each dimension could be used to evaluate ToM competences at any age (specifically, those related to desire, desire action, standard and false belief, and one of the indicators of mental-physical distinction); other composite variables were effective only among the youngest children (specifically, tasks related to basic emotions such as happiness, fear, sadness, and tasks involving the distinction between mental and physical entities when perceiving things that could be seen or touched), while other composite variables were effective only with older children (specifically, tasks related to surprise, one of the indicators of the mental-physical distinction, tasks involving the distinction between mental and physical entities when perceiving things that could be smelt, and beliefs different from standard or false beliefs).

Thus, although our study was not designed to explore the continuous or discontinuous nature of children's developmental mind comprehension, our findings support the idea that ToM is a continuously developing competence, whose key and discriminating components change throughout early infancy and childhood: emotion and desire understanding are the salient abilities at preschool age, whereas beliefs and mental-physical understanding are the more salient skills at primary school age.

The current study is complementary to Molina and Bulgarelli (2012): it enlarges the standardization sample, it confirms the reliability characteristics of the test and it investigates the factorial structure which has not previously been analysed with an Italian sample. Future research

could focus on two aspects. First, our final model was guided by empirical data, and needs to be confirmed with independent samples. Second, both the PCM analysis and the CFA showed that some items and composite variables functioned differently according to children's age and these differences should be explored in greater depth.

In sum, the ToM Storybooks offers some interesting features. First, it was designed to cover a wide age range (3-8 years), allowing ToM development to be evaluated across preschool and early primary school ages. Second, differently from the ToM Tasks or the ToM Test, the ToM Storybooks has been standardized and validated in two different European countries. Last, but not least, the test allows reliable evaluation of both overall ToM and four of its components, across a wide age range: for this reason, it may become a useful tool for longitudinal studies examining the issue of continuity versus discontinuity. In fact, as briefly discussed earlier, children's performance on an implicit location-false-belief task has been found to predict their later scores on a explicit location task but not on an explicit content-false-belief task (Thoermer et al, 2012), that is to say, ToM ability has been found to display continuity between similar tasks but not across different tasks. Thus, exploration of the issue of continuity in ToM should take a range of ToM components into account in order to verify whether continuity regards general competence in ToM or only specific components of it, and the ToM Storybooks may be used for this purpose.. Moreover, in clinical practice, the ToM Storybooks is a resource that may be used to build up differential profiles for specific populations. For instance, ToM competence can vary greatly among children affected by Pervasive Developmental Disorder and the ToM Storybooks may be of help in identifying which areas of ToM currently present deficits, in order to define more focused intervention.

References

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: New developments and techniques.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Baron-Cohen, S. (2000). Theory of mind and autism: A review. *International review of research in mental retardation*, *23*, 169–184. doi: 10.1016/S0074-7750(00)80010-5

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37–46. doi: 10.1016/0010-0277(85)90022-8

Blijd-Hoogewys, E. M. A., van Geert, P. L. C., Serra, M., & Minderaa, R. B. (2008). Measuring Theory of Mind in Children: Psychometric Properties of the ToM Storybooks. *Journal of Autism and Developmental Disorders*, *38,* 1907–1930. doi: 10.1007/s10803-008-0585-3

Bulgarelli, D., Molina, P., & Salomone, E. (2011). Valutare la competenza emotiva: significato e strumenti nello sviluppo tipico e atipico [Evaluation of emotional competence: meaning and tools in typical and atypical development], in: I. Grazzani & C. Riva Crugnola (Ed.) *Lo sviluppo della competenza emotiva: percorsi tipici e atipici*, Milano, Unicopli, pp. 203–216.

Bulgarelli, D., Blijd-Hoogewys, E. M. A., van Geert, P. L. C., & Molina, P. (2011, August). *Developmental trends in Theory of Mind understanding: data from two European countries*. Communication at the XV European Conference on Developmental Psychology, Bergen (N).

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337-342. doi: 10.1111/j.1467-8624.1996.tb01822.x

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., ... & Singh, S. (2005). Synchrony in the onset of mental-state reasoning evidence from five cultures. *Psychological Science*, *16*(5), 378-384. doi: 10.1111/j.0956-7976.2005.01544.x

Caron, A. J. (2009). Comprehension of the representational mind in infancy. *Developmental Review*, *29*(2), 69-95. doi: 10.1016/j.dr.2009.04.002

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive development*, *9*(4), 377-395. doi: 10.1016/0885-2014(94)90012-4

Flavell, J. H., Miller. P. H., & Miller, S. A. (1993). *Cognitive Development*. Hemel Hempstead: Prentice Hall.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi: 10.1080/10705519909540118

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. doi: 10.1111/1469-7610.00633

Istat (2003). *Famiglia, abitazione e luogo in cui si vive. Indagine multiscopo sulle famiglie "Aspetti della vita quotidiana" – anno 2002* [Family, home and place. Multi-purpose survey on families "Everyday characteristics" – year 2002] [Adobe Digital Editions version]. Retrieved from http://www3.istat.it/dati/catalogo/20040116_00/

Istat (2006). *Strutture famigliari e opinioni su famiglia e figli. Indagine multiscopo sulle famiglie "Famiglia e soggetti sociali" – anno 2003* [Family structure and opinions about family and offspring. Multi-purpose survey on families "Family and social subjects" – year 2003] [Adobe Digital Editions version]. Retrieved from http://www3.istat.it/dati/catalogo/20040116_00/

Lecce, S., & Hughes, C. (2010). The Italian job?: Comparing theory of mind performance in British and Italian children. *British Journal of Developmental Psychology*, *28*, 747–766. doi: 10.1348/026151009X479006

Linacre, J. M. (2009) *Winsteps* (Version 3.68.0) [Computer Software]. Chicago: Winsteps.com.

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: state of the art.*British Journal of Developmental Psychology*, *30*(1), 1-13. doi: 10.1111/j.2044-835X.2011.02074.x

Mayer, A., & Träuble, B. E. (2013). Synchrony in the onset of mental state understanding across cultures? A study among children in Samoa. *International Journal of Behavioral Development*, *37*(1), 21-28. doi: 10.1111/j.0956-7976.2005.01544.x

Molina, P. & Bulgarelli, D. (2012). La standardizzazione italiana del test ToM Storybooks: dati preliminari [Italian standardization of the ToM Storybooks: preliminary data]. *Giornale Italiano di Psicologia*, *39*(4), 863–879. doi: 10.1421/73146

Molina P., Bulgarelli D., Henning A., e Aschersleben G. (2014). Emotion undestanding: A cross-cultural comparison between Italian and German preschoolers. *European Journal of Developmental Psychology*. doi: 10.1080/17405629.2014.890585

Mitchell, P. (1997) *Introduction to theory of mind. Children, autism and apes*. London: Arnold.

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. doi: 10.1023/A:1025922717020

Muthén, L. K., & Muthén, B. O. (2004). *MPLUS user's guide*. Los Angeles, CA: Muthén & Muthén.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?. *Science*, *308*(5719), 255-258. doi: 10.1126/science.1107621

Sodian, B. (2011). Theory of mind in infancy. *Child Development Perspectives*, *5*(1), 39-43. doi: 10.1111/j.1750-8606.2010.00152.x

Steerneman & P., Meesters, C. (2009), ToM test-R. Handleiding. Antwerpen/Apeldoorn: Garant Uitgevers.

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology*, *30*(1), 30-44. doi: 10.1111/j.2044-835X.2011.02046.x

Tager-Flusberg, H., & Sullivan, K. (2000). A componential view of theory of mind: Evidence from Williams syndrome. *Cognition*, *76*(1), 59–90. doi: 10.1016/S0010-0277(00)00069-X

Thelen, E., & Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, Mass.: The MIT Press.

Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology*, *30*(1), 172-187. doi: 10.1111/j.2044-835X.2011.02067.x

Thommen, E., Cartier-Nelles, A., Guidoux A., & Wiesendanger, S. (2011). Aspects typiques et atypiques du développement de la théorie de l'esprit [Typical and atypical developmental characteristics of Theory of Mind]. *Approche Neuropsychologique des Apprentissages chez l'Enfant*, *112-113*, 145–151.

van Geert, P. L. C. (2003). Dynamic Systems Approaches and Modeling of Developmental Processes. In: J. Valsiner, & K. J. Conolly (Eds.), *Handbook of Developmental Psychology* (pp. 640-672). London: Sage.

Vesterinen, J. (2008). *The ToM Storybooks as a tool of studying children's Theory of Mind in Finland* (Master's thesis). Università di Jyväskylä, Jyväskylä, SF.

Vinden, P. G. (1996). Junin Quechua children's understanding of mind. *Child Development*, *67*(4), 1707-1716. doi: 10.1111/j.1467-8624.1996.tb01822.x

Vinden, P. G. (1999). Children's understanding of mind and emotion: A multi-culture study. *Cognition & Emotion*, *13*(1), 19-48. doi: 10.1080/026999399379357

Wellman, H.M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Wellman, H. M. (2012). Theory of mind: Better methods, clearer findings, more development. *European Journal of Developmental Psychology*, *9*(3), 313-330. doi: 10.1080/17405629.2012.680297

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. Child development, 72(3), 655–684. doi: 10.1111/1467-8624.00304

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential Progressions in a Theory-of-Mind Scale: Longitudinal Perspectives. *Child development*, *82*(3), 780-792. doi: 10.1111/j.1467-8624.2011.01583.x/

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541. doi: 10.1111/j.1467-8624.2004.00691.x

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding and deception. *Cognition*, *13*, 103–128. doi: 10.1016/0010-0277(83)90004-5

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief?. *British Journal of Developmental Psychology*,*30*(1), 156-171. doi: 10.1111/j.2044-835X.2011.02060.x

Appendix

List of the theoretical composite variables used in Blijd-Hoogewys *et al* (2008), briefly described with the relative abbreviation and number of items (dichotomous + justifications).

	Composite variables description	Abbreviation	n. items
1.	Recognition of basic emotions, namely happiness, sadness, anger, fear and surprise	Emotion Recognition (ER)	5+0
2.	Naming of basic emotions, namely happiness, sadness, anger, fear and surprise	Emotion Naming (EN)	5+2
3.	Recognition of emotions (happiness and surprise) produced by beliefs, for example: Sam wants something, he thinks it is in a specific place, he finds it, therefore he is happy	Beliefs Emotion Recognition (BER)	2+0
4.	Naming of emotions (happiness and surprise)	Beliefs Emotion Naming (BEN)	2+1
5.	Recognition of emotions (happiness and sadness) generated by desires, e.g., Sam wants something, he does not obtain it, therefore he is sad	Desire Emotion Recognition (DER)	5+1
6.	Naming emotions (happiness and sadness) generated by desires	Desire Emotion Naming (DEN)	5
7.	Appreciating the role of desire in generating behaviours, e.g. Sam wants something, he goes searching for it	Desire Action (DA)	3+1
8.	Distinguishing between mental entities, which are private and intangible, and physical entities, which are public and perceivable via different senses (seeing, touching, hearing)	Mental - Physical senses (MP senses)	8+2
9.	Distinguishing between mental entities, which are private and intangible, and physical entities, which are public and perceivable by different subjects	Mental - Physical others (MP others)	4+1
10.	Distinguishing between mental entities, which are private and intangible, and physical entities, which are public and perceivable by different times (present vs. future events)	Mental-Physical future (MP future)	4+1
11.	The ability to recognize that some physical phenomenona called Close Impostors (smoke, sounds) may be perceived with some senses only: (seing or smelling vs. touching)	Close Impostors (CP senses)	4+0
12.	The ability to recognize that some physical phenomenona called Close Impostors (smoke, sounds) may be perceived similarly by different persons	Close Impostors (CP others)	2+1
13.	The ability to recognize that some physical phenomenona called Close Impostors (smoke, sounds) may be perceived in the present but not in the future	Close Impostors (CP future)	2+1
14.	Distinguishing between real and unreal entities (thoughts and dreams)	Real and imaginary (RI)	7+0
15.	Appreciating the role of standard belief in generating behaviours	Standard Belief (SB)	3+1
16.	Classical false belief and explicit false belief tasks	False Belief (FB)	5+2
17.	Understanding that if someone's belief changes, their behaviour will be coherent with their current belief	Changed Belief (CB)	1+1
18.	Appreciating the role of standard belief in generating behaviour when the sentence is expressed in negative form	Non-Belief (NB)	2+1
19.	Understanding that story characters behave coherently with their beliefs even when these are contrary to those of the child taking the test	Not Own Belief (NOB)	1+1
20.	Role of deductive processes in generating belief (inferred belief control)	Inferred Belief (IB)	4+0
21.	Awareness that perception is essential to acquiring knowledge of the physical world	Seeing is KnoWing (SKW)	1+1

Table 1

Goodness of fit indices for the two confirmatory factor models (N=206 age 3-4; N=243 age 5-6; N=232 age 7-8)

Model	n. of composite variables	χ^2MLR	df	p	RMSEA (CI 90%)	CFI	SRMR
Initial model age 3-4	24	430.3	243	< .001	.061 (.052-.071)	.77	.08
Initial model age 5-6	22	279.5	200	< .001	.040 (.028-.051)	.88	.06
Initial model age 7-8	19	274.4	143	< .001	.063 (.052-.074)	.65	.07
Final model age 3-4	17	117.9	110	.286	.019 (.000-.041)	.99	.05
Final model age 5-6	21	245.7	180	< .001	.039 (.026-.050)	.90	.06
Final model age 7-8	12	67.8	48	< .05	.042 (.013-.064)	.92	.05

Note: RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Residuals

Table 2

Standardized loadings of the final models in the three age groups (age group 1/age group 2/age group 3)

Dutch composite variables (Blijd-Hoogewys et al., 2008)	PCM composite variables (current study)	n.	Factors				
			Emotion	Desire	Mental-Physical	Belief	Perception Knowledge
Emotion (E): naming, recognition	E Anger	1	.52/.46 /-				
	E Fear	2	.41/.37/-				
	E Happiness	3	.59 /.56 /*				
	E Sadness	4	.58/.52 /*				
	E Surprise	5	.24 /.27/.44				
Belief emotion (BE): naming, recognition	BE happiness	6	.45/.41 /*				
	BE surprise	7	-/.27/.68				
Desire emotion (DE): recognition, naming	DE Happiness task 1	8	.55/.36 /-				
	DE Happiness task 2	9	.49/.51/-				
	DE Sadness task 1	10		.60/.67/.54			
	DE Sadness task 2	11		.70/.61/.38			
	DE Sadness task 3	12		.50/.64/.43			
Desire action (DA)	DA	13				.59 /.49/.32	
Mental physical (MP): senses, others and future	MP Task 1 & 3	14			.63/.69/.53		
	MP Task 2 & 4	15			-/.47/.46		
Close impostor (CI): senses, others and future	CI smelling	16			-/.44/.55		
	CI Seeing	17			-/.28 /-		
	CI Touching	18			.71/.54 /-		
Real-Imaginary (RI)	RI Real exists	19			-/ */ *		
	RI Unreal exists	20			.74/ */*		
	RI Think about real/unreal	21			-/ - /.43		
Other belief tasks (B1)	CB, NB, NOB, IB	22				-/.26/.58	
Standard belief and False belief (B2)	SB, FB	23				.58/.63/.53	
Seeing is knowing	Seeing is knowing	24					1.0/1.0/x

p< .005 for all the loadings; Composite variable excluded from models due to: ceiling effect (*); loadings with wrong sign or not statistically significant at 0.05 (-); correlation (standardized covariance) with the others ToM factors not statistically significant (x)

Table 3

Factors correlations (standardized covariances) of the final models in the three age groups (age group 1/age group 2/age group 3)

	Emotion	Desire	Mental-Physical	Belief	Perception knowledge
Emotion	-				
Desire	.49/.33/ .33	-			
Mental-Physical	.51/.63/ .66	.11[a]/.17[a]/.20[a]	-		
Belief	.63/.71/ .40	.30/.35/.08[a]	.60/.88/.94	-	-
Perception knowledge	.26/.44/ x	-.01[a]/.22/ x	.32/.46/ x	.51/.47 / x	-

$p < .05$ for all the correlations except (a); dimension non included in the model (x)
Emotion (Emotion naming and recognition; belief-emotion), Desire (Desire predicting sadness), Mental-Physical (distintio between mental and physical phenomena), Belief (standard and non standard belief, false belief), Perception knowledge (seeing is knowing)

Table 4
Mean and (standard deviation) of the ToM Storybook CFA dimensions for age semesters; p value of the linear trend and Eta squared (η^2) in the last columns.

	Age group 1				Age group 2					Age group 3				
	3 (N=53)	3.5 (N=86)	4 (N=59)	p /η^2	4.5 (N=52)	5.0 (N=57)	5.5 (N=72)	6.0 (N=62)	p /η^2	6.5 (N=73)	7.0 (N=58)	7.5 (N=39)	8.0 (N=61)	p /η^2
Emotion	5.8 (3.7)	8.4 (4.3)	9.3 (4.4)	< .001 .096	13.6 (5.0)	14.9 (4.1)	16.1 (3.2)	17.4 (3.4)	< .001 .114	2.8 (1.6)	3.0 (1.7)	2.7 (1.8)	3.5 (1.6)	n.s. .033
Desire	1.5 (1.9)	2.0 (1.9)	2.9 (2.2)	< .001 .067	3.4 (2.2)	3.5 (2.2)	3.6 (2.3)	3.9 (2.0)	n.s. .008	3.8 (1.8)	3.8 (2.0)	4.1 (1.7)	4.3 (1.7)	n.s. .014
Mental-Physical	7.8 (3.4)	9.2 (3.3)	10.4 (3.2)	< .001 .081	18.5 (4.8)	19.3 (4.9)	20.8 (5.1)	22.1 (4.7)	< .001 .073	21.2 (4.0)	21.0 (3.5)	22.9 (3.0)	23.1 (4.1)	< .001 .065
Belief	6.2 (2.1)	8.0 (3.6)	8.2 (3.4)	< .001 .066	14.0 (4.5)	14.7 (3.9)	15.1 (4.7)	16.7 (4.3)	< .001 .046	17.4 (3.8)	17.7 (4.4)	19.0 (3.7)	19.4 (4.2)	< .001 .045
Perception Knowledge	0.5 (0.7)	0.7 (1.0)	0.7 (0.9)	n.s. .007	1.6 (1.3)	1.8 (1.3)	2.2 (1.1)	2.0 (1.1)	< .05 .036					