J Autism Dev Disord
DOI 10.1007/s10803-011-1244-7

ORIGINAL PAPER

Psychometric Evaluation of the Theory of Mind Inventory (ToMI): A Study of Typically Developing Children and Children with Autism Spectrum Disorder

Tiffany L. Hutchins · Patricia A. Prelock · Laura Bonazinga

Abstract Two studies examined the psychometric properties of the Theory of Mind Inventory (ToMI). In Study One, 135 caregivers completed the ToMI for children (ages 3 through 17) with autism spectrum disorder (ASD). Findings revealed excellent test–retest reliability and internal consistency. Principle Components Analysis revealed three subscales related to the complexity of ToM understanding. In Study Two, data were collected for 124 typically developing children (2 through 12 years). Findings again revealed excellent test–retest and internal consistency. The ToMI distinguished groups by age (younger vs. older children) and developmental status (typically developing vs. ASD), and predicted child performance on a ToM task battery. Utility of the ToMI, study limitations and directions for future research are discussed.

Keywords Autism · Theory of Mind · Validity · Measurement

Introduction

Theory of Mind (ToM) has been used, more or less, interchangeably with terms like 'perspective-taking', 'social cognition', 'metacognition', and 'folk psychology' (Astington and Baird 2005; Flavell et al. 2002). It has also been described in broad terms as "a conceptual system that underlies our folk psychology with which we impute mental states to others and ourselves" (Perner and Lang 2000, p. 150) and "an intentional stance that characterizes human social interaction" (Astington 2003, p. 14). Indeed, ToM has been described as a term that "refuses to be corralled" (Astington and Baird 2005, p. 4) and so one useful description of ToM is its construal as a complex and multifaceted construct that reflects the understanding of an interconnected network of mental states.

Despite the tremendous activity in the field and developments in ToM assessment procedures (for reviews see Baron-Cohen 2000; Sprung 2010), the measurement of ToM has traditionally been associated with several challenges. First, standard ToM tasks (e.g., the Sally-Ann task and other structured elicitation procedures) are often scored on a dichotomous pass/fail basis and even ToM task batteries designed to assess a range of ToM competencies are limited in their ability to adequately tap the content domain relevant to construct of ToM (Hutchins et al. 2008a). In addition, it has long been recognized that the administration of many ToM tasks are complicated by cognitive and linguistic (e.g., Wellman et al. 2001) as well as situational (Tager-Flusberg 1999) and motivational (Begeer et al. 2003) performance factors. ToM tasks also often suffer from ceiling effects when mind-reading skills are relatively good (Slaughter and Repacholi 2003). This prompted the development of several 'advanced' tests of mind-reading that require implicit social interpretation (e.g., Baron-Cohen et al. 2001; Beaumont and Sofronoff 2008; Happé 1994) and the result is advancement in methods to assess ToM among individuals with high functioning autism and Asperger syndrome. On the other end of the spectrum, however, ''many people with autism cannot even be tested with standard theory of mind tasks, since they lack the cognitive and verbal skills necessary to answer the control

T. L. Hutchins (✉) · P. A. Prelock
Department of Communication Sciences and Disorders, University of Vermont, 489 Main Street, 407 Pomeroy Hall, Burlington, VT 05452, USA
e-mail: tiffany.hutchins@uvm.edu

L. Bonazinga
Burlington, VT, USA

Published online: 12 April 2011

questions, success on which is usually an inclusion criterion'' (Happé 1995, pp. 845–847). Thus, children with ASD who are nonverbal or have limited language skills are commonly excluded from research protocols. Finally, despite the findings that ToM tasks can correlate with observations of social functioning (e.g., Frith et al. 1994; Lalonde and Chandler 1995) the explicit nature of many traditional ToM tasks are quite unlike the ways that real life social dilemmas are presented (Astington 2003; Klin et al. 2000). In fact, the notion that ToM task performance can outstrip social cognitive functioning as it is applied in everyday interactions has been the topic of considerable concern (Astington 2003; Davies and Stone 2003; Klin et al. 2000).

In our initial effort to develop a ToM assessment that would address the aforementioned challenges, we described the development and psychometric evaluation of the *Perceptions of Children's Theory of Mind Measure-Experimental version* (PCToMM-E; Hutchins et al. 2008a, b). The measure was designed to serve as an index of caregivers' *perceptions* of children's ToM knowledge and, by proxy, children's *actual* ToM knowledge. The PCToMM-E consisted of 33 statements designed to tap a range of ToM competencies that varied in content and complexity.

In response to the breadth of the construct and guided by the immense literature on ToM in both typically developing children and children with ASD, we sought to develop face valid indicators including (but not limited to) child knowledge of, or ability to engage in pretence, desire and intentionality, distinctions between appearance and reality, causes of emotions, mental–physical distinctions, knowledge that seeing leads to knowing, first- and second order thinking, visual perspective-taking, affective recognition, empathy, and social and logical inferencing (Hutchins et al. 2008a, b).

The PCToMM-E was found to have excellent psychometric properties (e.g., test–retest reliability, convergent validity, contrasting-groups construct validity) when administered to caregivers of typically developing children (ages 2–12) and caregivers of children with ASD (ages 2–12). No ceiling effects were observed for the sample with ASD and ceiling effects for the typically developing sample did not occur until late childhood (i.e., around age 10 or 11). We concluded that the PCToMM-E demonstrated good evidence for construct validity and that it was sensitive to fine variations in ToM across the age ranges and linguistic skills represented. In that report, we also examined parents' ability to accurately predict the performance of their children with ASD (who varied widely in verbal ability) on a ToM task battery. We found that parents of children with ASD are excellent and reliable sources of information about their children's ToM knowledge. We also argued that the use of parents as informants, who are uniquely situated to observe their child's ToM as demonstrated (or not demonstrated) during real world social interaction, helped to move us toward an assessment of ToM that is socially valid and family-centered. Having established a degree of confidence that the PCToMM-E could play an important role in ToM assessment, we began to focus our attention on how to improve the measure.

Limitations of the PCToMM-E

Although the PCToMM-E addressed some of the problems raised in the literature regarding the breadth and relevance of ToM assessment, it was not without limitations in sampling, age ranges used, and item clarity. These limitations, discussed in greater detail below, provided an opportunity to consider changes that would enhance its utility as a research and assessment tool.

The sample used to validate the PCToMM-E involved only 20 caregivers of children with ASD and 60 caregivers of typically developing children. Although this sample provided sufficient power to reveal significant effects using inferential comparisons, the sample was not adequate for factor analysis. Factor analysis is an important next step in the validation of the measure. It allows for an examination of the dimensionality of the measure and the generation of subscale scores reflecting values of the underlying constructs for use in additional analyses.

Moreover, the original measure was developed using individuals with ASD up to age 12 and there was no evidence of ceiling effects even at the oldest ages sampled. Because only modest improvements in ToM knowledge have been observed in adolescence (e.g., Frith et al. 1994), we reasoned that our measure would not evidence ceiling effects when extended to older individuals with ASD. Since confirmation of this assumption is important for enhancing the tool's research and clinical utility, the revised measure was evaluated for individuals with ASD through age 17.

Some items on the original version were revised for clarity or omitted based on improvement in Cronbach's alpha. Several items to tap more complex or advanced aspects of ToM were also added with hopes of raising the ceiling scores for typically developing children. These included the ability to understand humor (St. James and Tager-Flusberg 1994), sarcasm (Baron-Cohen 1997), counterfactual reasoning (Perner et al. 2004) and the distinction between jokes and lies (Leekam and Prior 1994) and knowing and guessing (Flavell 1999). Items tapping the understanding of the mind as an active interpreter (e.g., understanding biased interpretation and understanding that two people can see the same image and interpret it differently) were also included as these are considered advanced ToM competencies (Pillow 1991).

The continuum and hash mark response arrangement (see "Appendix") was retained in the new version of the measure because it offers respondents the ability to indicate degrees of certainty between anchors and is more sensitive than a numeral and Likert-type scale in that regard. Because this response arrangement is relatively uncommon, however, confusion over how to indicate a response can occur. We noticed that respondents sometimes indicated their response by circling the anchors or by making an 'X' along the continuum. To enhance clarity, we added information to our instructions and provided examples of incorrect and correct response forms. In addition, when respondents felt complete confidence in their child's acquisition of a particular ToM competency, they were instructed to indicate this by making a hash mark through the circle at the end of the continuum.

The revised measure (now referred to as the Theory of Mind Inventory; ToMI) was evaluated for two different populations in two separate studies. Study one (study two is described later) investigated the psychometric properties of the ToMI when completed by primary caregivers of children with ASD. First, the test–retest reliability of the ToMI was examined. The measure's criterion-related construct validity was then explored by evaluating the strength of the association between scores on the ToMI and verbal mental age (VMA) because VMA has been implicated as a strong predictor of ToM abilities in children with ASD (e.g., Happé 1995). We also expected that a construct valid index of ToM would yield scores that would correlate with children's performance on a ToM task battery. Finally, exploratory principle components analysis was performed to examine the dimensionality of the measure.

Study 1: Method

Participants

A national sample of 104 ToMI forms were completed by primary caregivers (99 mothers and 5 fathers) of children diagnosed with ASD. The national sample was drawn from 14 states and each major geographic region in the United States (i.e., Northeast, Southeast, Northwest, Southwest, Midwest). A local (Vermont) sample of 31 mothers and their children diagnosed with ASD also participated in this study.

For the combined sample ($n = 135$), caregiver age ranged from 24 to 59 ($M = 42$; SD = 6.8), education ranged from 12 (completion of high school) to 20 (doctoral degree) years ($M = 15.40$; SD = 2.21), and gross annual combined income ranged from less than \$1,000–\$4,00,000 ($M =$ \$67,350; SD = \$48,283). All caregivers identified as the child's primary caregiver and reported spending an average of 7.8 h with the child per day (SD = 3.38; not counting when the child was sleeping). A total of 127 (94.1%) caregivers were native English speakers, all were fluent in English, and 19 (14.1%) were fluent in more than one language.

Children were 27 females (20%) and 108 males (80%) who ranged in age from 3 years, 4 months to 17 years, 8 months ($M = 10.12$, SD = 4.19). On the basis of caregiver report, 74 (55%) children were currently diagnosed with autistic disorder, 30 (22.2%) were diagnosed with Pervasive Developmental Disability—Not Otherwise Specified (PDD-NOS), 29 (21.5%) were diagnosed with Asperger Syndrome, and 2 reported "other" (and indicated "autism spectrum disorder"). Also according to parent report, 20 (14.8%) children were characterized as nonverbal, 24 (17.7%) were characterized as having limited language, and 91 (67.5%) were characterized as verbal (see description of each category below).

Measures

Demographic Questionnaire

A demographic questionnaire was administered to all participants (i.e., local and national samples). The questionnaire solicited relevant demographic information including, but not limited to, caregivers' and child's date of birth and gender, parent education, annual combined household income, and information related to child grade level, diagnosis, and language level. With regard to the assessment of language level, we adapted an item from the Vineland Adaptive Behavior Scale (Sparrow et al. 2005) which is a measure of adaptive behavior from birth to adulthood which is commonly used in the assessment of individuals with ASD. The item asked respondents to indicate whether the child's language level was best characterized as nonverbal (no functional speech), limited speech (uses meaningful 2–3 word phrases on a regular basis), or verbal (uses language in a variety of ways to communicate with others).

The Theory of Mind Inventory (ToMI)

The revised measure (now referred to as the *Theory of Mind Inventory*; ToMI) consisted of 48 statements accompanied by a response continuum of 20 metric units (equivalent to 6.75 inches using a 30 ft/inch engineering scale) anchored by 'definitely not' and 'definitely' with a center point of 'undecided' (see "Appendix" for an example of the continuum). Respondents are instructed to carefully read each statement and indicate their degree of confidence that each statement is true or not true by placing a hash mark at the appropriate point along the continuum. Responses for each item are scored by a ruler (possible range = 0–20) and averaged with higher values reflecting greater degrees of certainty that the target child possesses ToM knowledge across the range of content surveyed. The ToMI now yields both composite and subscale scores which were derived on the basis of principle components analysis (described below) using data obtained for our ASD sample.

Peabody Picture Vocabulary Test-4

The *Peabody Picture Vocabulary Test* (PPVT-4; Dunn and Dunn 2007) is a measure of receptive vocabulary, which has gained considerable currency as an index of verbal ability in the ASD literature. The test requires identification of visual stimuli, which increase in difficulty. The PPVT-4 has been evaluated extensively for reliability (internal consistency, alternate forms, test–retest) and validity (content, construct, convergent) and has been found to have excellent psychometric properties (Shaw 2008).

Theory of Mind (ToM) Task Battery

The *ToM Task Battery* (Hutchins et al. 2008a) originally consisted of 16 test questions within nine tasks designed to tap a range of ToM tasks. The items range in complexity and are presented in the form of static visual stimuli. The initial task tests the ability to identify emotions associated with facial expressions. The second task requires the child to understand the visual perspective of the examiner. The third, fourth and fifth tasks measure the child's ability to infer desire-based emotion, perception-based belief, and perception-based action, respectively. The remaining four tasks assess advanced abilities including first- and second-order false beliefs. The ToM Task Battery was developed in response to the need for a ToM battery that would be appropriate for verbal and nonverbal children with ASD. The test was constructed so that children could respond by answering verbally or by pointing to a picture that showed the correct answer. Each ToM question is scored as pass (1) or fail (0). In line with previous research, a child does not receive credit for a ToM target question without passing the associated control questions.

The *ToM Task Battery* has been evaluated for reliability (test–retest, internal consistency; Hutchins et al. 2008a). Thirteen of the original 16 items achieved adequate or excellent reliability. Of the three items that were not found to be reliable, one was dropped from the battery and the remaining two items were revised (Hutchins et al. 2008a).

Procedure

National Sample

The national sample was recruited through professional contacts of the second author who were engaged in service delivery for children with ASD and their families. An IRB approved e-mail was sent to each contact explaining the purpose of the research and a request to distribute the demographic questionnaire and ToMI to families with children with ASD. Professional contacts who responded affirmatively shared these materials with families who were asked to return the forms in a prepaid self-addressed envelope.

Local Sample

The local sample was recruited through an invitation to parents and children who had previously participated in our ongoing research program focusing on interventions to remediate the core deficits of ASD (described in Hutchins and Prelock 2006, 2008). Other participants were recruited via contact letters from local support agencies for families of children with ASD. On the basis of parents' reports, a developmental pediatrician or psychologist with experience in autism diagnosed all the children with autism ($n = 19$), pervasive developmental disorder-not otherwise specified ($n = 9$), or Asperger syndrome ($n = 3$).

All data collection procedures were conducted in the families' homes during a single visit that lasted approximately 1 h. First, mothers completed the demographic questionnaire and the ToMI. Then the children were administered the PPVT-4 and the ToM task battery. For the purposes of test–retest reliability, all respondents were provided an additional copy of the ToMI and asked to complete it between 2 and 6 weeks later. Respondents who complied dated and returned this second copy in a self-addressed stamped envelope.

Study 1: Results

Descriptive Statistics for the ToMI

For the combined sample, ToMI scores ranged from 1.1 to 19.95 ($M = 10.8$; SD = 4.5). Descriptive data by age (year) are provided in Fig. 1 for the entire sample. Because some comparisons were conducted on the basis of the child's verbal abilities, descriptive data for this variable are also offered. The average score for the combined sample for children characterized as nonverbal on the basis of parent report was 6.54 (SD = .71), which was similar to the average score for children with limited language

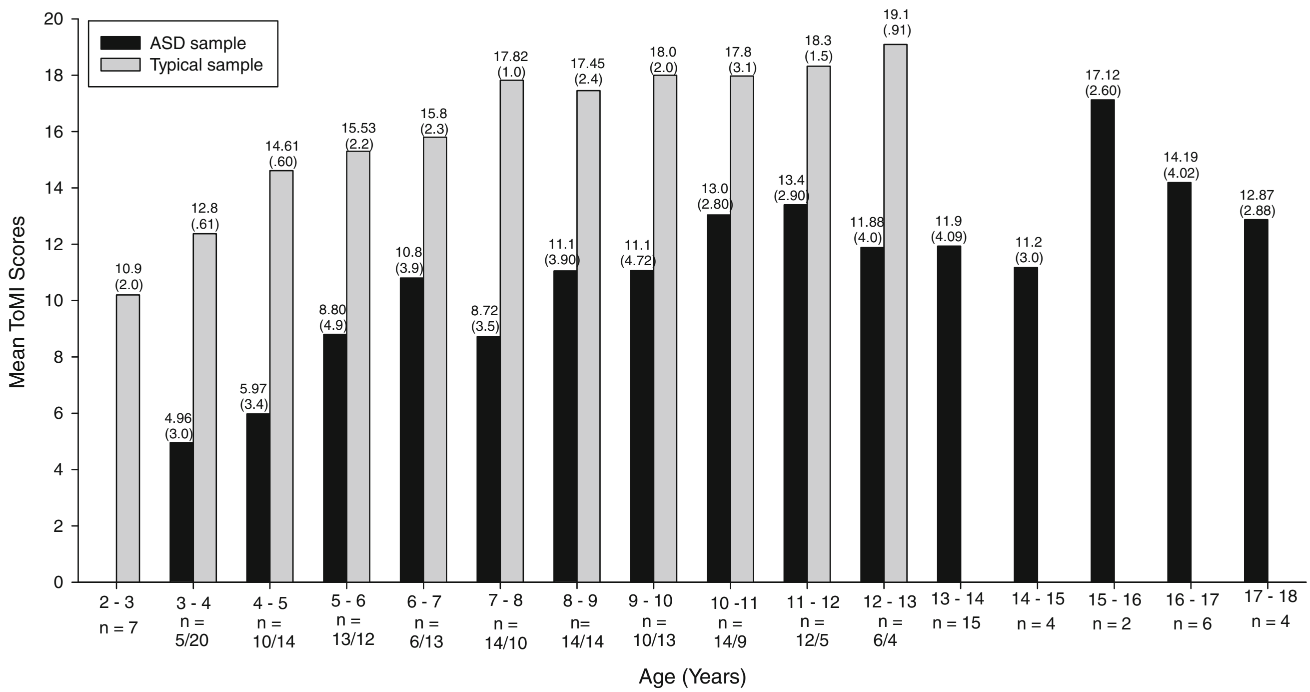


Fig. 1 Descriptive data by age for total ToMI scores for ASD and typical sample: Mean (SD)

($M = 6.52$; SD = .70). The average score for children characterized as verbal was 12.68 (SD = .35). With regard to receptive language (local sample only; as measured by the PPVT-4), 11 children scored three standard deviations below the mean, five scored two standard deviations below the mean, and two scored one standard deviation below the mean. An additional 10 children scored in the normal range, two scored one standard deviation above the mean, and one scored three standard deviations above the mean. These data indicate that children with ASD with a wide range of verbal abilities were represented in this study.

Reliability

Test–Retest Reliability

Test–retest data were missing for two respondents who failed to complete the ToMI at time two. Thus, data for 29 parents from the local sample who completed the ToMI at time two provided estimates of test–retest reliability. A Pearson's product moment correlation indicated strong stability of this measure using an interval of 14–78 days (M = 27.5; SD = 19.4; $r = .89$, $p < .001$) which is a highly dependable relationship with variation in scores at time 1 accounting for approximately 80% of the variation in scores at time 2. The SEM was 1.50 which is acceptable (i.e., it is less than one-third of the standard deviation; McCauley 2001).

Internal Consistency

Internal consistency was assessed using Cronbach's alpha which resulted in a very high estimate of consistency of content ($\alpha = .98$). Examination of alpha based on item deletion indicated that removal of two items resulted in slight improvement in Cronbach's alpha. These two items ("My child understands the word desire" and "My child understands the word need" which appeared to assess knowledge of low frequency mental state terms) were subsequently dropped from further analysis.

Validity

Criterion-Related Validity

As expected, a Pearson's product moment correlation indicated a substantial positive relation between ToMI scores and PPVT-4 standard scores ($r = .73$, $p < .05$) with variation in ToMI scores accounting for approximately 53% of the variation in children's receptive vocabulary.

A construct valid measure of ToM competence should also be positively correlated with children's scores on ToM tasks. A Spearman's rho (because ToM task battery data are best construed as ordinal in nature) indicated a substantial positive relationship ($r = .66$, $p < .05$) with variation in scores on the ToMI explaining approximately 44 percent of the variation in children's scores on the ToM task battery.

Principal Components Analysis

Data for all participants (national and local sample, $n = 135$) were submitted to exploratory principal components analysis (PCA). Three rotation methods were explored: one was orthogonal (varimax) and two were oblique (direct oblimin and promax). Although oblique rotation is recommended for correlated factors (which applied to some of our factors), all rotation methods resulted in nearly identical factor solutions and varimax rotation was most effective in achieving a simple structure. In fact, whether factors are correlated may not make much difference in the exploratory stages of analysis. "It even can be argued that employing a method of orthogonal rotation (or maintaining the arbitrary imposition that the factors remain orthogonal) may be preferred over oblique rotation, if for no other reason than that the former is much simpler to understand and interpret" (Kim and Mueller 1978, p. 50). Of course, many measures remained complex in that they continued to load on more than a single factor (>.30).

Varimax rotation with Kaiser normalization yielded a six factor solution. Upon inspection of the rotated matrix, factors 4 and 5 were comprised of two items and were dropped from further analysis as is convention (Costello and Osborne 2005). Factor 6 included one item (item 2) which was moved to Factor 1 where it had its second highest loading and a strong theoretical basis. Factors 1–3 accounted for 62.7% of the cumulative variance (i.e., factor 1 = 52% with Eigenvalue of 24.9; factor 2 = 6.4% with Eigenvalue of 3; factor 3 = 4.4% with Eigenvalue of 2.1). The rotated component matrix for the three retained factors are presented in Table 1. The descriptive data for average subscale scores (i.e., factors 1–3) by age are presented in Fig. 2.

Study 1: Discussion

Descriptive data (even for older participants) reveal that the composite and subscale scores did not approach the ceiling. These findings extend the use of the ToMI as a valuable parent-report measure of ToM. First, many traditional ToM tests and even ToM task batteries evidence ceiling effects which limits their utility when mentalizing is relatively good. Descriptive data also repeated the finding that ToM remains impaired at comparable levels among adolescents with ASDs and that variability in ToM may increase over time (Peterson et al. 2005). Of course, it must be noted that the sample size for a few ages (e.g., age 15 years) are still small and should not be assumed to yield stable descriptive statistics at this time. The ToMI also evidenced excellent test–retest and internal reliability suggesting good temporal stability and consistency of content.

Although there is disagreement as to the details, for example, the age of attainment of a particular concept and the causes of development, "there is general agreement among theorists of theory of mind about the general pattern of development" (Flavell et al. 2002, p. 189). This general agreement is seen in the typically developing literature as ToM competencies are linked to chronological age. By contrast, ToM development in the ASD literature is not tied to chronological timetables but it is similar in that it has been distinguished along dimensions of relatively early emerging, simple, or foundational skills to relatively late emerging, complex, or advanced skills (Hutchins and Prelock 2008).

As described below, each factor for the ASD sample looked, first and foremost, like a complex of interrelated competencies reflecting a *general* level in the typical developmental progression of ToM. We do not intend that these results be taken as evidence that the developmental progression of (or mechanisms operating in) ToM in typical children and individuals with ASD are equivalent and evidence to the contrary is accumulating (e.g., Peterson et al. 2005). We do suggest that the items comprising each factor hang together because their expression in day-to-day functioning, as observed reliably by caregivers, share some underlying cognitive features or capacities.

Taken in reverse order, the content of factor 3 seems most characteristic of the earliest ToM competencies that are known to emerge in typical development during infancy and toddlerhood. For example, the ability to engage in joint or shared attention (intended to be tapped by items 43 and 44) is estimated to emerge around 9 months of age (Tomasello 1995) and has been described as a foundational skill of cosmic importance for the advancement of social cognition (Bates 1979; Carpendale and Lewis 2006). Another ToM development associated with 9 months is social referencing (item 33) or the tendency of infants to look to a parent when faced with ambiguous events (Walden and Ogan 1988). Of course, social referencing may involve the ability to interpret others' facial expressions; an ability that we attempted to tap using at least two other items that also loaded on this factor (items 7, 30). Late toddlerhood has also been credited (although not without controversy) as a time when children can demonstrate some understanding of others' intention including whether acts are purposeful or accidental (item 28; Astington 1991).

Clearly, researchers disagree about whether a 'rich' (i.e., epistemic) or 'lean' (i.e., non-epistemic) interpretation of early ToM competencies is best supported by the available evidence (e.g., Carpendale and Lewis 2006; Flavell et al. 2002) and the present study is not designed to inform this question. The important point here is that, on their face, the items that comprised factor 3 appear to represent some of

Table 1 Rotated factor solution for three retained factors (loadings < .3 are suppressed for ease of interpretation)

	Factor 1	Factor 2	Factor 3
Factor 1 (n = 16)			
2. If it were raining and I said in a sarcastic voice "Gee, looks like a really nice day outside," my child would understand that I didn't actually think it was a nice day	**.39**		
6. My child understands that people can be wrong about what other people want	**.59**		
16. If I said "Let's hit the road!" my child would understand that I really meant "Let's go!"	**.53**	.35	
17. My child understands that people can lie to purposely mislead others	**.63**	.51	
20. My child understands that people can smile even when they are not happy	**.74**		
21. My child understands the difference between when a friend is teasing in a nice way and when a bully is making fun of someone in a mean way	**.73**		
22. My child understands that people don't always say what they are thinking because they don't want to hurt others' feelings	**.84**		
23. My child understands the difference between lies and jokes	**.69**		
24. My child understands that if two people look at the same object from a different standing point, they will see the object in different ways	**.74**		
26. My child understands that people often have thoughts about other peoples' *thoughts*	**.75**	.35	
27. My child understands that people often have thoughts about other peoples' *feelings*	**.66**	.42	
32. My child recognizes when a listener is not interested	**.56**		.51
39. My child is able to put himself/herself in other people's shoes and understand how they feel	**.68**		
42. If I said "What is black, white and '*read*' all over? It's a newspaper!" my child would understand the humor in this play on words	**.66**	.41	
46. When we like others, we are likely to interpret their behavior in positive ways and when we don't like others, we are likely to interpret their behavior more negatively. My child understands that previous ideas and/or opinions of others can influence how we interpret their behaviors	**.60**		
47. My child understands that two people can see the same image and interpret it differently. For example, when looking at this image, one person might see a rabbit whereas another might see a duck	**.52**	.46	
Factor 2 (n = 19)			
1. My child understands that when someone puts on a jacket, it is probably because he/she is cold		**.56**	
4. My child understands that when someone says they are afraid of the dark, they will not want to go into a dark room		**.65**	.35
8. My child understands the word 'think'	.44	**.62**	
10. If I put my keys on the table, left the room, and my child moved the keys from the table to a drawer, my child would understand that when I returned, I would first look for my keys where I left them	.39	**.44**	
12. My child understands that to know what is in an unmarked box, you have to see or hear about what is in that box		**.47**	
13. My child understands the word 'know'		**.71**	
14. Appearances can be deceiving. For example, when seeing a candle shaped like an apple, some people first assume that the object is an apple. Given the chance to examine it more closely, people typically change their mind and decide that the object is actually a candle. If my child was in this situation, my child would understand that it was not the object that changed, but rather his or her ideas about the object that changed	.35	**.54**	
15. If I showed my child a cereal box filled with cookies and asked "What would someone who has not looked inside think is in the box?", my child would say that another person would think that there was cereal in the box		**.72**	
18. My child understands that when someone makes a 'guess' it means they are less certain than when they 'know' something	.56	**.62**	
19. My child understands that when someone is thinking about a cookie, they cannot actually smell, eat or share that cookie	.49	**.64**	
31. My child can pretend that one object is a different object (for example, pretending a banana is a telephone)		**.55**	.52
34. My child understands the word 'if' when it is used hypothetically as in, "*If I had the money, I'd buy a new house*."	.41	**.56**	
35. My child understands that when a person uses his/her hands as a bird, that the person doesn't actually think it is a real bird		**.71**	
36. My child knows how to make up stories to get what he/she wants	.43	**.56**	
37. My child understands that in a game of hide and seek, you don't want the person who is 'it' to see you		**.79**	
38. My child understands that when a person promises something, it means the person is supposed to do it		**.71**	
40. My child understands that when someone shares a secret, you are not supposed to tell anyone	.43	**.67**	

Table 1 continued

	Factor 1	Factor 2	Factor 3
45. My child understands the word 'believe'	.53	**.65**	
48. My child understands that if Bruce is a mean boy and John is a nice boy, Bruce is more likely than John to engage in malicious or hurtful behaviors		**.68**	
Factor 3 (n = 7)			
3. My child recognizes when someone needs help			**.54**
7. My child understands that when people frown, they feel differently than when they smile			**.56**
28. My child understands whether someone hurts another on purpose or by accident	.41		**.58**
30. My child recognizes when others are happy			**.62**
33. My child understands that, when I show fear, the situation is unsafe or dangerous		.32	**.67**
43. My child is able to show me things		.43	**.64**
44. My child is able to pay attention when I show him/her something			**.72**

Bolded values reflect the highest factor loadings

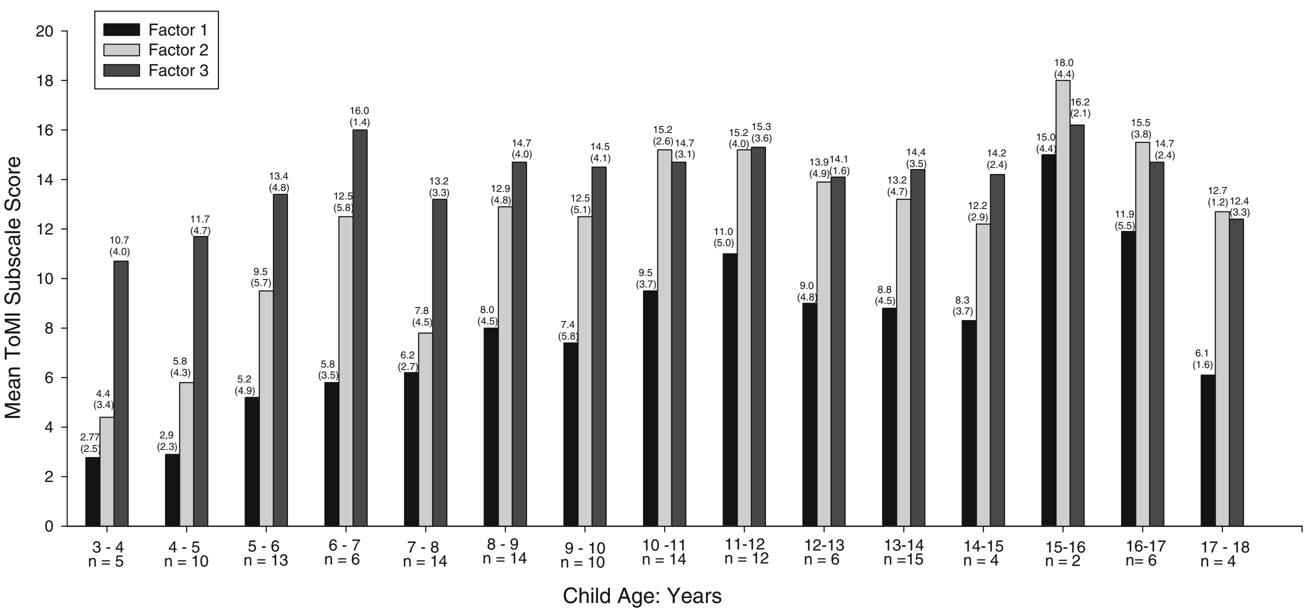


Fig. 2 Descriptive data by age and factor (subscale) scores for ASD sample: Mean (SD)

the most significant ToM achievements that have been implicated (although variably interpreted) in infancy and toddlerhood. This factor was termed "Early ToM: Reading Affect and Sharing Attention."

By contrast, factor 2 items seemed to tap ToM advancements characteristic of typically developing preschool children. Many of these are believed to require a basic metarepresentational skill with which a child not only makes use of mental representations, but also knows that these are, in fact, representations. By the age of approximately 4-years (because this depends on how cognitively and linguistically demanding the task is), typically developing children can demonstrate an understanding of metarepresentation. This ability is associated with pretence (item 31; Leslie 1987) and is seen as underlying success on tasks involving false belief (items 10 and 15), the appearance-reality distinction (item 14), the mental-physical distinction (items 19 and 35), and seeing-leads-to-knowing (items 12 and 37), to name a few (e.g., Baron-Cohen 1989; Gopnik and Astington 1988; Wimmer and Perner 1983). Metarepresentation is further implicated in the understanding of how psychological states guide behavior (items 1, 4; e.g., Leslie and Frith 1998) as well as attempts to engage in deception (item 36; e.g., Perner 1991) although these attempts may vary considerably in their sophistication. Items intended to tap the understanding (*not* the earlier production) of cognitive terms (items 8, 13, 18, and 45) also loaded on factor 2 which has been reported to emerge around age 4 years (Abbeduto and Rosenbberg 1984; Kazak et al. 1997; Moore and Furrow 1991) and is reinforced by the finding that "proper comprehension of the lexical items 'know' and 'think' tends to go together with correct prediction of behavior" (Leslie and Frith 1998, p. 322). The understanding of speech acts (items 38 and 40) and counterfactual reasoning (item 34) also loaded on factor 2. For each of these, the meta-level of understanding has been seen as crucial although a mature understanding of these aspects tend to emerge later than 4 years (Astington 1988; Perner et al. 2004). This factor was termed "Basic ToM: Metarepresentation and Developmentally Related Understandings."

Metarepresentation is commonly viewed as a prerequisite skill that is necessary, but not sufficient, for an implicitly held theory of mind (Astington 2003; Frith and Frith 2000). Factor 1 items appeared to tap the most advanced ToM knowledge intended to be captured by the measure, most of which are described in the literature as emerging in typical development between the ages of 6 and 8 years. These items appear to require complex recursion, metapragmatic and metalinguistic skills, and an understanding of the mind as an active interpreter.

Recursive thinking requires the embedding of representations (e.g., Tiffany thinks about what Patty thinks) "and so first order false belief understanding is one example. However, there are more complex types of recursion than false belief understanding and these form important aspects of human thought about social matters" (Carpendale and Lewis 2006, p. 190). These more complex forms of recursion include second-order beliefs (items 6, 26 and 27) as well the distinction between lies and jokes as this requires both the understanding that a falsehood is intended and the understanding of whether the falsehood is intended to be believed (item 23; Leekam and Prior 1994; Sullivan et al. 1995). Sophisticated metapragmatic competencies include the socially appropriate use of display rules (items 20 and 22; Flavell et al. 2002) so that, for example, someone might smile even though she is unhappy. Advanced metalinguistic understanding involves not only the ability to identify the listener's belief and the speaker's intention but to distinguish various types of speech acts from each other (Keenan 2003). This understanding includes, but is not limited to, knowledge of nonliteral idiomatic uses (item 16; Keenan 2003), sarcasm (item 2; Keenan 2003), and humor (e.g., play on words, item 42; McGhee 1979).

Other items loading on factor 1 were designed to tap biased cognition (item 46) which is the understanding that one's previous experiences or expectations about others may color their interpretation of events (Pillow 1991). Another item was intended to tap knowledge of the mind as an active interpreter (item 47; Carpendale and Chandler 1996). Indeed, more advanced ToM competencies "include a commonsense understanding that knowledge is interpretive and that the mind itself influences how the world is experienced" (Carpendale and Lewis 2006, p. 193).

A mature ToM also involves the ability to make accurate social judgments (items 21 and 32) which is a particularly advanced skill (and is particularly difficult for individuals with ASD) considering that all social encounters are embedded in context. Among other things, it involves reading mental states and attitudes that may be revealed in subtle social cues and understanding their relation to the physical and social environment to extract meaningful and relevant information. In sum, the aforementioned advanced aspects of social cognition not only underscore the complexity and multifaceted nature of ToM, they also remind us that there is more to a mature understanding of ToM than mastery of false beliefs (and metarepresentation more generally) that emerge later in typical course of development. This factor was termed "Advanced ToM: Complex Recursion, Mind as Active Interpreter, and Social Judgment."

Although these three general developmental levels (what we term Early, Basic, and Advanced) were revealed in our analyses, it is important to note that they should not be considered discrete or inflexible (or even fixed) stages. We agree with Chandler (1988) who defends the notion that "there are more intervening stages in this process than are generally supposed" (p. 389) and that the development of ToM is "considerably less homogenous and linear, and substantially more protracted and differentiated" (p. 390). In addition, the items were not constructed to tap discrete components of ToM functioning insofar as each is conceptualized as a complex aspect that undoubtedly has connections to others within the broad and multifaceted construct that is ToM. Accordingly, we not only assume variation in ToM across but also *within* factors. For example, there is some evidence to support the idea that pretence is a developmental precursor to false belief understanding (Leslie 1987) although the two items designed to tap these domains both loaded on factor 2.

With regard to the factor loadings, many items were complex (i.e., they loaded on more than one factor) and this may be partly explained by the interconnected nature of ToM dimensions described above. It might also be related to the nature of parents' interpretations of children's reasoning. If children exhibit a behavior that resembles what adults do, then it may be reasonable to attribute the same kinds of understanding. On the other hand, "the same behavior could be accompanied by different kinds of understanding" (Moore and Povinelle 2007, p. 216). Thus, it is possible that parents sometimes attribute adult-like reasoning to children whose social understandings are not that advanced (and the potential for underestimation of reasoning skills exists as well). We expect that parents differ greatly in their tendency to view their own understandings and the behavior-based-understandings they attribute to their children diagnosed with ASD as analogous. The degree to which parents were inaccurate in attributing more or less advanced reasoning on the ToMI is not known but it may, in part, explain some of the complex loadings that were revealed in our factor structure.

In a related vein, not all items comprising each factor demonstrated perfect alignment with the typically developing literature and a few factor loadings challenged this interpretation. As a case in point, we developed one item (item 24) to assess Level 2 visual perspective taking which credits both the understanding that something is seen as well as the appreciation that how something is seen can differ depending on perspective (Flavell 1992). This ability is strongly associated with metarepresentation and has been found to be present in children ages 4 to 5-years (Flavell 1992; Flavell et al. 2002). From the standpoint of a developmental progression, then, this item should have clustered with factor 2 items but instead it loaded with factor 1 which was interpreted to reflect more developmentally advanced capacities.

What might explain the performance of an item like the one designed to assess Level 2 visual perspective-taking? One possibility is error due to item construction. An alternative interpretation may reflect differences in the developmental progression of ToM among individuals with ASD (or the cognitive strategies they employ). Still another explanation recognizes the fact that parents are reporting on child competencies using information they have accumulated through observations of the child who endeavors to solve ToM problems as they are presented in real-world samples of behavior. This is important because, as noted previously, the ways that most ToM tasks are demonstrated by way of standard ToM tasks are often very different from how they are presented in real-life (Astington 2003; Davies and Stone 2003; Klin et al. 2000) and this raises issues related to the social validity of ToM measurement.

Study 2

A second study was conducted to evaluate the psychometric properties of the ToMI when completed by primary caregivers of typically developing children. Study Two examined the reliability (test–retest, internal consistency) and validity (criterion-related, contrasting groups) of the measure. Factor analysis was not conducted for the typical sample because the wide variation in developmental level results in ceiling effects for particular items at particular ages. Thus factor analysis for a group of typically developing children ranging from 2 to 12 would not be expected to result in a coherent set of latent variables.

Recall that the PCToMM-E evidenced ceiling effects in late childhood. Specifically, scores for typically developing children on the original measure approached the ceiling around age 9 or 10 and reached it around age 11 years. Several items were included in the ToMI in an effort to tap more complex or advanced aspects of ToM. We expected this to depress scores generally and hoped it would result in retreat from the ceiling for all children as well as enhanced differentiation across the ages sampled.

Study 2: Method

Participants

A national sample of 37 ToMI forms were completed by primary caregivers (all mothers) of children identified as typically developing. Unlike the data for the ASD national sample discussed above, the sample for typically developing children was limited to only five states (i.e., Vermont, Texas, Illinois, California, and Massachusetts). A local (Vermont only) sample of 60 mothers of 87 typically developing children also participated in this study.

For the combined samples ($n = 124$), caregiver age ranged from 23 to 52 ($M = 38$; SD = 5), education ranged from 12 (completion of high school) to 20 (doctoral degree) years ($M = 16.67$; SD = 1.98), and gross annual combined income ranged from less than \$20,000–\$2,75,000 ($M =$ \$102,355; SD = \$44,210). All caregivers identified as the child's primary caregiver and reported spending an average of 7.4 h with the child per day (SD = 2.46; not counting when the child was sleeping). A total of 120 (97%) caregivers were native English speakers, all were fluent in English, and 16 (12.9%) were fluent in more than one language.

Children were 62 females (50%) and 62 males (50%) who ranged in age from 2 years, 0 months to 12 years, 8 months ($M = 7$, SD = 2.22). On the basis of caregiver report, no child had ever been diagnosed with a disability (including uncorrected visual or hearing impairment) with the exception that 4 had received services in the past to remediate a speech (not language) impairment.

Procedure

National Sample

The sample recruitment procedures were identical to those described above with the exception that caregivers of typically developing children were targeted.

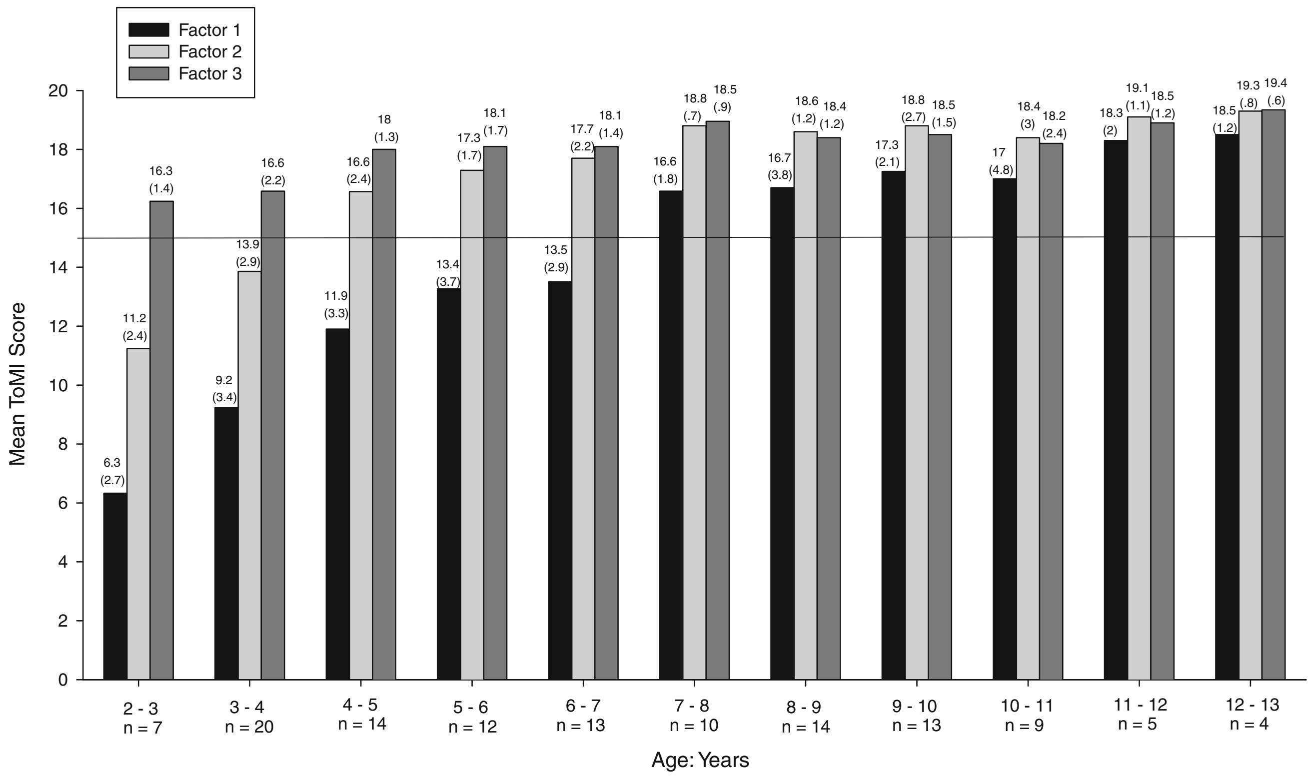


Fig. 3 Descriptive data by age and factor (subscale) scores for typical sample: Mean (SD)

Local Sample

Participants (60 mothers of 87 typically developing children) were recruited by circulating a recruitment letter to local preschools and elementary schools and to faculty in the psychology and communication sciences departments at a local university as well as through informal personal contacts. All data collection procedures took place in the participants' homes during a single visit that lasted approximately 1 h. First, the demographic questionnaire and the ToMI were completed. To evaluate the criterion-related validity of the ToMI when administered to typically developing samples, children (ages 2–9 only to avoid ceiling effects) were then administered the ToM task battery. For the purposes of test–retest reliability, all respondents were provided an additional copy of the ToMI and asked to complete it between 2 and 6 weeks later. Respondents who complied dated and returned this second copy in a self-addressed stamped envelope.

Study 2: Results

Descriptive Statistics for the ToMI

For the combined sample, ToMI scores ranged from 4 to 20 ($M = 15.62$; SD = 3.34). Descriptive data for the overall ToMI scores were examined by age. These data are presented in Fig. 1. Inspection of the descriptive data by subscale scores were also informative. These data are presented in Fig. 3. A line corresponding to a score of 15 has been added to the graph to indicate the point at which parents, on average, endorsed ToM competencies as "probably" present (discussed below).

Reliability

Test–Retest Reliability

Twenty-eight mothers (46.6% response rate) of 37 children completed and returned the measure at time two to provide estimates of test–retest reliability. The test–retest interval was between 12 and 44 days ($M = 25.46$; SD = 10.43). Data for the 37 responses indicated high test–retest reliability ($r = .89$, $p < .01$). This is a highly dependent relationship with variation in scores at time one explaining approximately 80% of the variation in scores at time two. The SEM was 1.11 which is acceptable (again, it was less than one-third of the standard deviation).

Internal Consistency

Internal consistency was assessed using Cronbach's alpha which resulted in a very high estimate of content consistency (α = .98). As with the ASD sample, examination of alpha based on item deletion indicated that removal of two items resulted in slight improvement in Cronbach's alpha. These two items (the same as indicated earlier) were dropped from further analysis.

Validity

Contrasting-Groups Developmental Method of Construct Validity

The age of four years is often considered a time when significant ToM developments can be observed. In a meta-analysis of 178 studies employing a false belief task, Wellman et al. (2001) found that children under age 3.5 years typically perform below chance, children between the age of 3.5 and 4 years typically perform at chance, and children four and older typically perform above chance.

Data from the combined sample of typically developing 2.5–3.5 years olds ($n = 16$) were compared to data from mothers of typically developing 4–5 year olds ($n = 13$). An independent t-test indicated that there was a significant difference between mothers of 2.5–3.5 year olds ($M = 11.27$; SD = 2.40) and mothers of 4–5 years olds ($M = 14.79$; SD = 2.08), $t\ (27) = 4.12$, $p < .01$.

We also expected a construct valid measure of ToM to distinguish typically developing children and individuals with ASD. An independent t test revealed a significant difference, $t\ (257) = 10.04$, $p < .001$, such that mothers of children identified as having ASD (ages 3–17) reported lower scores ($M = 10.8$; SD = 4.42) than did mothers of younger (ages 2–12) typically developing children ($M = 15.6$; SD = 3.37).

Criterion: Related Validity

ToM not only develops in early childhood but it continues to develop into late childhood and beyond (although these studies are relatively rare). Thus, it was expected that a construct valid measure of ToM would show increases in scores that correspond to increases in child age among children who are typically developing (ages 2–12). A significant correlation was found between child age and ToMI scores ($r = .72$, $p < .05$). Variation in child age accounted for approximately 52% of the variation in ToMI scores.

Criterion-related validity was further examined by correlating ToMI and ToM task battery scores. A subset of 24 children (between 2 and 9 years) from the local sample completed the ToM Task Battery. Child participants between the ages of 2 and 9 years were chosen to avoid ceiling effects in ToM task battery performance. A Spearman's rho revealed a significant relationship ($r = .82$, $p < .05$) with variation in ToMI scores accounting for 67% of the variance in scores on the ToM Task Battery.

Study 2: Discussion

Compared to the original measure (Hutchins et al. 2008a, b), scores on the ToMI were lower (anywhere from 1 to 3 points depending on age sampled) and this was in line with our expectation that the inclusion of advanced items would depress scores (and raise ceilings). This was especially important to enhance the utility of the measure for the oldest typically developing children sampled. It is noteworthy that the Advanced subscale (i.e., factor 3) scores evidenced the greatest distinction across the older ages and so examination of scores for this subscale may be particularly useful in investigations of mentalizing in middle and late childhood. Nevertheless, the overall shape of the distribution of scores by age (see Fig. 2) is similar to one reported for the original measure (Hutchins et al. 2008a, b) and neither the composite score nor the Advanced subscale currently distinguish between older children who are contiguous in chronological year (e.g., scores do not distinguish between ages 7 vs. 8 and between ages 9 vs. 10). Although research in this area is limited, the addition of items that reflect competencies that demonstrate development over this period would be helpful for addressing this limitation.

Because the subscales were derived from analyses of the ASD sample and interpreted in light of typical developmental timetables, inspection of the average subscale scores by age when applied to the typically developing sample was warranted. If our interpretation of the factor structure was accurate, then the general levels of ToM that we identified (i.e., Early, Basic, Advanced) should be demonstrated in the scores of typical children as developing at a particular age. To establish a criterion for mastery, we examined subscale scores in relation to an average score of 15 (which corresponds to a parental endorsement of "Probably"). In line with our interpretation, inspection of Fig. 3 shows that scores for the Early ToM subscale (i.e., factor 3) met this criterion for the youngest ages sampled (2 years) which is consistent with the notion that these competencies are typically observed in infancy and toddlerhood. In addition, scores for the Basic ToM subscale (i.e., factor 2) met this criterion between the ages of 4 and 5 years which supports our interpretation that these competencies are related to the achievement of metarepresentation which is typically observed in preschool age children. Finally, scores for the Advanced ToM subscale (i.e., factor 1) met this criterion at age 7 which reinforces our interpretation that these skills represented more mature ToM skills that develop in later childhood. These descriptive data provide additional support for validity of the ToMI and its newly developed subscales.

The ToMI demonstrated excellent temporal and internal consistency as well as evidence for contrasting groups and criterion-related construct validity. With regard to validity, the ToMI distinguished groups on the basis of age (i.e., younger and older typically developing children) and developmental status (i.e., typically developing children from individuals with ASD). ToMI scores also correlated with age which was expected given that ToM continued to develop beyond the formative years. ToMI scores also predicted child performance on a ToM task battery.

General Discussion

As noted in this paper's introduction, several 'advanced' tests of ToM have been developed and have been useful in the study of individuals with high functioning autism and Asperger syndrome (Baron-Cohen et al. 2001; Beaumont and Sofronoff 2008; Happé 1994). On the other hand,

> "interest in people with high-functioning autism can obscure the fact that most people with the disorder have moderate to severe learning difficulties. In classic autism this may be about 75%, and more than half of those affected develop no appreciable language. This means that theory of mind deficits in autism have only been examined in a fraction of sufferers; typically experiments include only children with verbal mental ages of above 4 years" (Doherty 2009, p.179).

Clearly, there is a need for measures that can evaluate ToM among individuals with ASD with the most limited language and functioning. Further, a tool that is useful for individuals who vary widely in language and social cognitive functioning would facilitate the study of ToM across these populations. Finally, a tool that is not plagued by ceiling effects and that could be used appropriately for younger and older typically developing children *and* individuals with ASD offers an alternative, or complementary, method for assessing ToM for these between group comparisons.

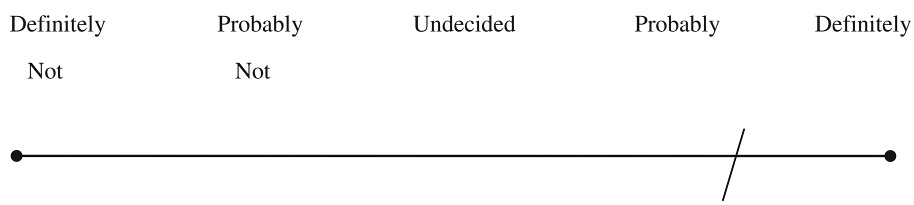
Of course, several limitations of this study and directions for future research deserve mention. First, although individuals with ASD were surveyed through age 17, the sample sizes for some of the older ages were small, so larger samples are needed to develop confidence in the stability of scores for some of the ages sampled (e.g., age 15). In a related vein, given the absence of a ceiling for the ASD sample, future efforts should evaluate the ToMI on an even wider age range of individuals with ASD. We expect ToMI scores would not approach the ceiling for younger and older adults with ASD and extending examinations of validity to these samples would enhance the utility of the measure. Because a primary aim of the ToMI was to develop a measure that would be appropriate for verbal and nonverbal individuals, a larger sample of individuals with ASD with the most limited language skills is desirable as the current sample was weighted toward those with functional language. Although the PPVT has gained some currency in the ASD literature as a means for indexing language skills, the PPVT is a measure of receptive vocabulary only and the use comprehensive language measures to better assess the intralinguistic profiles of the sample is needed.

Lerner et al. (2011) explored the relationship of ToMI scores to parent-reported autistic symptoms, social skills, and autism-related social impairments in a sample of adolescents with ASD. They found that ToMI scores were correlated with social skills, ASD symptoms, and autism related social impairments in the expected directions. In line with this approach, we are including a wider range of measures (i.e., those intended to tap social skills, social responsiveness, autistic symptoms, pragmatic language use, nonverbal intelligence, expressive vocabulary) for the purposes of additional criterion-related analyses. Future research should also examine whether and to what degree ToMI scores predict social functioning in real-world observations of behavior as this has implications for social validity. Going forward, it will be important to include tests of discriminant validity. The development of a multi-trait multi-method matrix to distinguish variance related to construct versus method of measurement (e.g., parent-report vs. child performance measures) would also be desirable.

The ToMI performed well in all examinations of reliability and validity in both typically developing children and individuals with ASD. The advantages of the ToMI over traditional ToM assessment procedures include the fact that responses are not scored on a dichotomous pass/fail basis but rather in terms of continuum of confidence that a child possesses a ToM understanding. A range of scores can be obtained which are construed as interval in nature and can be submitted appropriately to the most powerful statistical procedures. In addition, the ToMI is quick and easy to administer and it is not complicated by situational factors and child motivation, linguistic, or cognitive performance factors. We will be exploring methods for electronic scoring of the measure which would be particularly important for large research programs to expedite the scoring process. Finally, the measure now yields both composite and subscale scores that reflect general levels of ToM development. This may make the ToMI attractive as a clinical tool and we are currently exploring how the tool is best used to identify intervention targets and guide treatment planning. At present, this study provides strong support for the further development and evaluation of the ToMI for use in research and clinical settings.

Appendix

The continuum response arrangement used in the ToMI.



References

Abbeduto, L., & Rosenbberg, S. (1984). Children's knowledge of the presuppositions of know and other cognitive terms. *Journal of Child Language, 12*(3), 621–641.

Astington, J. W. (1988). Promises: Words or deeds? *First Language, 8*, 259–270.

Astington, J. W. (1991). Intention in the child's theory of mind. In D. Frye & C. Moore (Eds.), *Children's theories of mind: Mental states and social understanding* (pp. 157–172). Hillsdale, NJ: Erlbaum.

Astington, J. W. (2003). Sometimes necessary, never sufficient: False belief understanding and social competence. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind: Implications for typical and atypical development* (pp. 13–38). New York: Psychology Press.

Astington, J. W., & Baird, J. A. (2005). Introduction: Why language matters. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 3–25). New York: Oxford University Press.

Baron-Cohen, S. (1989). Are autistic children 'behaviorists'? An examination of their mental-physical and appearance-reality distinctions. *Journal of Autism and Developmental Disorders, 19*(4), 579–600.

Baron-Cohen, S. (1997). Hey! It was just a joke! Understanding propositions and propositional attitudes by normally developing children and children with autism. *Israel Journal of Psychiatry, 34*, 174–178.

Baron-Cohen, S. (2000). Theory of mind in autism: A fifteen year review. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from developmental neuroscience* (pp. 3–20). Oxford, UK: Oxford University Press.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The 'Reading the Mind in the Eyes' test revised version: A study of normal adults, and adults with Asperger syndrome or high functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241–251.

Bates, E. (1979). *The emergence of symbols: Cognition and communication in infancy*. New York: Academic Press.

Beaumont, R. B., & Sofronoff, K. (2008). A new computerized advanced theory of mind measure for children with Asperger syndrome: The ATOMIC. *Journal of Autism and Developmental Disorders, 38*(2), 249–260.

Begeer, S., Rieffe, C., Meerum Terwogt, M., & Stockmann, L. (2003). Theory of mind-based action in children from the autism spectrum. *Journal of Autism and Developmental Disorders, 33*, 479–487.

Carpendale, J. M., & Chandler, M. J. (1996). On the distinction between false belief understanding and subscribing to an interpretive theory of mind. *Child Development, 67*, 1686–1706.

Carpendale, J., & Lewis, C. (2006). *How children develop social understanding*. Malden, MA: Blackwell.

Chandler, M. (1988). Doubt and developing theories of mind. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing theories of mind* (pp. 387–413). New York: Cambridge University Press.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research and Evaluation, 10*(7), 1–9.

Davies, M., & Stone, T. (2003). Synthesis: Psychological understanding of social skills. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind: Implications for typical and atypical development* (pp. 305–352). New York: Psychology Press.

Doherty, M. J. (2009). *Theory of mind: How children understand others' thoughts and feelings*. New York: Psychology Press.

Dunn, L. M., & Dunn, L. M. (2007). *Examiner's manual for the peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Flavell, J. H. (1992). Perspectives on perspective-taking. In H. Beilin & P. Pufall (Eds.), *Piaget's theory: Prospects and possibilities*. Hillsdale, NJ: Erlbaum.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology, 50*, 21–45.

Flavell, J. H., Miller, P. H., & Miller, S. A. (2002). *Cognitive development* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Frith, C., & Frith, U. (2000). The physiological basis of theory of mind: Functional neuroimaging studies. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from developmental neuroscience* (pp. 334–356). Oxford, UK: Oxford University Press.

Frith, U., Happé, F., & Siddons, F. (1994). Autism and theory of mind in everyday life. *Social Development, 3*, 108–123.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development, 59*, 26–37.

Happé, F. (1994). An advanced test of theory of mind: Understanding the story characters thoughts and feelings by able autistic mentally handicapped and normal children and adults. *Journal of Autism and Developmental Disorders, 24*, 129–154.

Happé, F. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*, 843–855.

Hutchins, T., Bonazinga, L., Prelock, P. A., & Taylor, R. S. (2008a). Beyond false beliefs: The development and psychometric evaluation of the perceptions of children's theory of mind measure—experimental version (PCToMM-E). *Journal of Autism and Developmental Disorders, 38*(1), 143–155.

Hutchins, T. L., & Prelock, P. A. (2006). Using social stories and comic strip conversations to promote socially valid outcomes for children with autism. *Seminars in Speech and Language, 27*(1), 47–59.

Hutchins, T. L., & Prelock, P. A. (2008). Supporting theory of mind development: Considerations and recommendations for professionals providing services to individuals with autism spectrum disorder. *Topics in Language Disorders, 28*(4), 338–362.

Hutchins, T., Prelock, P. A., & Chace, W. (2008b). Test–retest reliability of theory of mind tasks representing a range of content and complexity and adapted to facilitate the performance of children with ASD. *Focus on Autism and Other Developmental Disabilities, 23*(4), 195–206.

James, P. J., & Tager-Flusberg, H. (1994). An observational study of humor in autism and Down syndrome. *Journal of Autism and Developmental Disorders, 24*(5), 603–617.

Kazak, S., Collins, G. M., & Lewis, V. (1997). Can young people with autism refer to knowledge states? Evidence from their understanding of 'know' and 'guess''. *Journal of Child Psychology and Psychiatry, 38*(8), 1001–1009.

Keenan, T. (2003). The preschool years and beyond. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind* (pp. 121–142). New York, Hove: Psychology Press.

Kim, J. O., & Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it*. Beverly Hills, CA: Sage.

Klin, A., Schultz, R., & Cohen, D. J. (2000). Theory of mind in action: Developmental perspectives on social neuroscience. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from developmental neuroscience* (pp. 357–390). Oxford, UK: Oxford University Press.

Lalonde, C. E., & Chandler, M. (1995). False belief understanding goes to school: On the social-emotional consequences of coming early or late to a first theory of mind. *Cognition and Emotion, 9*, 167–185.

Leekam, S. R., & Prior, M. (1994). Can autistic children distinguish lies from jokes? A second look at second-order belief attribution. *Journal of Children Psychology and Psychiatry, 35*(5), 901–915.

Lerner, M., Hutchins, T. L., & Prelock, P. A. (2011). Brief report: Preliminary psychometric evaluation of the ToMI and its relations to social skills. *Journal of Autism and Developmental Disabilities, 41*, 512–517.

Leslie, A. M. (1987). Pretense and representation: The origins of 'theory of mind'. *Psychological Review, 94*, 412–426.

Leslie, A. M., & Frith, U. (1998). Autistic children's understanding of seeing, knowing, and believing. *British Journal of Developmental Psychology, 6*, 315–324.

McCauley, R. (2001). *Assessment of language disorders in children*. Mahwah, NJ: Lawrence Erlbaum.

McGhee, P. E. (1979). *Humor: Its origins and development*. San Francisco: W. H. Freeman and Company.

Moore, C., & Furrow, D. (1991). The development of the language of belief: The expression of relative certainty. In D. Frye & C. Moore (Eds.), *Children's theories of mind: Mental states and social understanding* (pp. 173–193). Hillsdale, NJ: Erlbaum.

Moore, C., & Povinelle, D. J. (2007). Differences in how 12- and 24-month-olds interpret the gaze of adults. *Infancy, 11*(3), 215–231.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: Cambridge University Press.

Perner, J., & Lang, G. (2000). Theory of mind and executive function: Is there a developmental relationship? In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from developmental neuroscience* (pp. 150–181). Oxford, UK: Oxford University Press.

Perner, J., Sprung, M., & Steinkogler, B. (2004). Counterfactual conditionals and false belief: A developmental dissociation. *Cognitive Development, 19*, 179–201.

Peterson, C., Wellman, H., & Liu, D. (2005). Steps in theory of mind development for children with deafness or autism. *Child Development, 76*(2), 502–517.

Pillow, B. H. (1991). Children's understanding of biased social cognition. *Developmental Psychology, 27*(4), 539–551.

Shaw, S. R. (2008). Review of the *Peabody Picture Vocabulary Test* (4th ed). In K. F. Geisinger, R. A. Spies, J. F. Carlson & B. S. Plake (Eds.) The eighteenth mental measurements yearbook (pp. 143–144). Lincoln, NE: The Buros Institute of Mental Measurements.

Slaughter, V., & Repacholi, B. (2003). Introduction: Individual differences in theory of mind. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind* (pp. 1–12). New York: Taylor & Francis Books.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland adaptive behavior scales* (2nd ed.). San Antonio, TX: Pearson.

Sprung, M. (2010). Clinically relevant measures of children's theory of mind and knowledge about thinking: Non-standard and advanced measures. *Child and Adolescent Mental Health, 15*, 204–206.

Sullivan, K., Winner, E., & Hopfield, N. (1995). How children tell a lie from a joke: The role of second-order mental state attributions. *British Journal of Developmental Psychology, 13*, 191–204.

Tager-Flusberg, H. (1999). The challenge of studying language development in children with autism. In L. Menn & N. Bernstein Ratner (Eds.), *Methods for studying language production* (pp. 313–331). Mahwah, NJ: Lawrence Erlbaum.

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. J. Dunham (Eds.), *Joint attention: It's origins and role in development* (pp. 103–130). Hillsdale, NJ: Erlbaum.

Walden, T., & Ogan, T. (1988). The development of social referencing. *Child Development, 59*, 1230–1240.

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

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