The Validity and Scalability of the Theory-of-Mind Scale with Toddlers and Preschoolers

Rachel M Hiller, Nathan Weber & Robyn L Young

Flinders University, South Australia

Author Notes.

Rachel M Hiller, Nathan Weber, Robyn L Young: School of Psychology, Flinders University, South Australia, Australia.

Rachel M Hiller is now at Department of Psychology, University of Bath, Bath, United Kingdom.

The authors wish to thank Nicole Reid for her assistance with inter-rater reliability, as well as the Centres, Kindergartens and families who participated in this research.

Correspondence concerning this article should be addressed to Rachel Hiller, Department of Psychology, University of Bath, Claverton Down, Bath, BA2 7AY, United Kingdom. Email: R.Hiller@bath.ac.uk

The authors declare no conflicts of interest.

Despite the importance of theory of mind (ToM) for typical development, there remain two key issues affecting our ability to draw robust conclusions. One is the continued focus on false-belief as the sole measure of ToM. The second is the lack of empirically validated measures of ToM as a broad construct. Our key aim was to examine the validity and reliability of the five-item ToM scale (Peterson, Wellman, & Lui, 2005). In particular, we extended on previous research of this scale by assessing its scalability and validity for use with children from two years of age. Sixty-eight typically developing children (aged 24 to 61 months) were assessed on the scale's five tasks, along with a sixth Sally-Anne false-belief task. Our data replicated the scalability of the five tasks for a Rasch- but not Guttman-scale. Guttman analysis showed a four-item scale may be more suitable for this age range. Further, the tasks showed good internal consistency and validity for use with children as young as two years of age. Overall, the measure provides a valid and reliable tool for the assessment of ToM, and in particular the longitudinal assessment of this ability as a construct.

Theory of mind, toddlers, scale, validity

The past decade has seen a plethora of research assessing individuals' abilities to interpret the mental states of others, referred to as theory of mind (ToM). This construct represents an individual's ability to not only understand the perspective of others, but to interpret how this perspective may influence the person's behaviour. Literature on ToM ability has had a large impact in the fields of both developmental psychology and psychopathology. In particular, the extensive literature on ToM in late pre-school and early school years has consistently shown its importance for typical social development (Hughes & Leekam, 2004), including associations with peer rejection (Devine & Hughes, 2012), indirect aggression (Renouf et al., 2010), and prosocial behaviour (Caputi, Lecce, Pagnin, & Banerjee, 2012). Perhaps most notably, ToM is also considered a key cognitive deficit in individuals with autism spectrum disorder (Baron-Cohen, Tager-Flusberg, & Cohen, 2000).

Despite the importance of ToM for typical development, there remain two key limitations that affect our ability to draw robust conclusions on the developmental role of this construct. First, there is a lack of evidence regarding the validity and reliability of ToM measures, with some studies showing widely used ToM tasks have quite poor psychometric properties (Mayes, Klin, Tercyak, Cicchetti, & Cohen, 1996). The second concern is the literature's continued focus on false-belief ability as the sole measure of ToM (for similar critique see Burack, Charman, Yirmiya, & Zelazo, 2001). This focus contrasts with knowledge that ToM encompasses a range of abilities that begin to develop long before false-belief understanding (Wellmann, Cross, & Watson, 2001). Thus, the focus on a single ToM skill not only impedes our ability to understand ToM prior to four years of age, but also does not allow for the longitudinal assessment of ToM development. Although some measures have aimed to address the paucity of knowledge on early theory of mind, including measuring the gaze patterns of infants in a violation of expectation paradigm; Onishi & Baillargeon, 2005) and parent report on toddler's theory of mind abilities (theory of mind

inventory; TOMI; Hutchins, Prelock & Bonazinga, 2011), there remains a paucity of research on how children's explicit theory of mind may be directly assessed.

In an attempt to provide a more comprehensive ToM assessment, Wellman and Liu (2004) proposed a ToM scale, consisting of five tasks designed to assess first-order ToM abilities. The abilities measured were proposed to progress in difficulty and, as such, develop in a progression across early childhood. Wellman and Liu proposed that a single score calculated from the scale could be used to index ToM ability. That is, from a given ability level it could be presumed all lower items had been answered correctly and all higher (more difficult) items were answered incorrectly. The authors' original analysis of 75 typically developing children showed 80% of respondents (aged 3.5 years to 6 years of age) provided response patterns where this index score (highest item scored correct) did indeed accurately reflect their performance. The usefulness of this task has been further demonstrated with both typically- and atypically-developing children across numerous age ranges from three up to thirteen years old (Peterson, et al., 2005; Peterson, Wellman, & Slaughter, 2012; Shahaeian, Peterson, Slaughter, & Wellman, 2011).

Despite the scale's promise, some important questions remain regarding both its practical application and psychometric properties. To address these questions this study had three aims. First, the scale is proposed to provide a measure which assesses ToM skills that precede false-belief understanding. However, there is no evidence for the usability of the scale with children under 36 months of age, with the majority of research focussed on children over 42 months of age (when we would expect false-belief ability to emerge). As such, the first aims of this research was to assess the usability and scalability of the five items with children from two years of age. Our third aim was to assess the validity of this test against a standard false-belief task (the Sally-Anne paradigm). Through the assessment of this scale with a younger age range, we have provided information on the usability of this

scale as a single comprehensive test of early theory of mind, paving the way for its use as a longitudinal measure.

Method

Participants and Procedure

This study received ethical approval from the Flinders University Social and Behavioural Research Ethics Committee and the local Department of Education and Child Services (DECS) ethics committee. Information sheets were sent to parents at two local childcare centres and two local kindergartens. Parents were required to provide consent for the child's participation in the study. Participants were excluded if there was a suspected or known developmental delay. Parental consent was provided for 70 children. Of these children, two were unable to participate due to one child being ill and the second receiving a diagnosis of Autistic Disorder during the assessment period. The 68 remaining children (boys = 41, girls = 27) were aged between 24 and 61 months (M = 44.81, SD = 10.82). Of these children 19 were between 2 years 0 months and 2 years 11 months, 19 were between 3 years 0 months and 9 years 11 months, while 30 were between 4 years 0 months and 5 years 0 months of age (with majority under 4.5 years old). All children participated in individual cognitive assessments in a quiet area of their childcare centre or kindergarten. All tasks were administered on a small table by the primary researcher. When at the table the researcher sat directly opposite the child and to maintain motivation each child received a sticker following their participation in each task. No child indicated they wished to cease participation in the tasks, which were all play-like in nature and as such, all tasks were completed in one sitting for all participants.[1]

[1] During the recruitment period the researcher spent some time building rapport (albeit superficial) with the children. This did not specifically target participating children but rather involved the researcher being present for common group activities such as 'group-time' and lunch time. This ensured that the researcher was a 'familiar face' to the children in the centres, so the participating children were comfortable to leave the group to participate in the cognitive assessment.

Tasks and Scoring

Two separate measures were administered. One was the widely used single-item Sally-Anne false-belief task (see Baron-Cohen, Leslie, & Frith, 1985) and the other was the Australian adaptation of the five-item ToM scale (Peterson, et al., 2005). It is understood that this task is suitable for less verbal children and uses vocabulary with which Australian children are more familiar (e.g., 'biscuit' instead of 'cookie'). The scale's five tasks were: (1) Diverse Desires (understanding another's likes may differ from your own), (2) Diverse Beliefs (understanding another person may think differently about the same situation), (3) Knowledge Access (ability to judge another's knowledge of a scenario), (4) False Belief (judging another's false-belief about the content of a descriptive box), and (5) Hidden Emotion (understanding a person's facial expression does not always match their emotion; see Wellman & Liu, (2004)). For all children the scale's items were delivered from easiest to most difficult.

Tasks from the ToM scale were administered and scored as they were by Peterson and colleagues (2005). For a more detailed description of the procedures replicated by this study see Peterson, et al., 2005. Briefly, each task comprised a control question and a focal question. The purpose of the control question was to assess the child's understanding of the scenario presented, while the focal question assessed theory of mind. As with previous use of the scale (e.g., Wellman & Liu, 2004), for the child to receive a score of correct for the task, they were required to first respond correctly to the control question.

All task responses (to both the control and focal questions) were coded as either correct (1) or incorrect (0). For tasks one through four, children were able to respond verbally (requiring a single-word unambiguous response, such as 'biscuit') or by pointing to a picture (see Peterson et al., 2005). Task 5 included an additional justification component, where the child was required to give a brief explanation for their response. Responses to the

justification component of the Hidden Emotion task were recorded verbatim for later coding. An independent rater coded Hidden Emotion responses for the entire sample, with 100% interrater agreement.

Results

Our primary aim was to assess the validity and scalability of the five-item ToM scale for use with children from two years of age. Thus, we first examined the descriptive statistics to determine if the tasks were actually comprehensible to our younger age range (see Table 1). Children demonstrated comprehension of the control question for each of the first three tasks (Diverse Desires, Diverse Beliefs and Knowledge Access) by around two years of age. Indeed, the entire sample was able to correctly respond to the control question presented with task 1. There was also evidence of children demonstrating an understanding of diverse desires and diverse beliefs ToM abilities from just after two years of age.[2]

Insert Table 1 here

Scale Analyses

We assessed the scalability of the tasks to determine the best index (i.e., highest item correct or total items correct) of ToM performance. To replicate the analyses used by Peterson and colleagues (2005) scalability was assessed using Guttman (Green, 1956) and Rasch (Rasch, 1960) scale analyses. There are six patterns of responses consistent with scalable performance on the five items (see Table 2). These patterns reflect the key requirement of a scalable measure – specifically, that from a person's ability level all higher tasks should be

[2] A gender difference was only evident on task 3 (Knowledge Access). After controlling for age, compared to those children who passed this task, failing the task was predictive of being a boy, $Wald(1) = 6.29$, $p = .01$, $B = -1.89$, $SE = .76$.

incorrect and all lower tasks correct. Seventy-two per cent of the children in our sample responded in a pattern that mapped onto one of the six ordered patterns. Table 2 shows the percentage of children whose responses fit each possible pattern. The 'other' category in Table 2 represents response patterns that did not fit one of the scalable patterns.

Insert Table 2 here

Of those children whose responses did not fit an exact scalable pattern, 74% (n = 14) gave an incorrect response followed by a correct response on the next highest item. This discrepancy did not appear to occur consistently for any specific item. The remaining five children (26%) from the 'other' category, received incorrect scores on two items below their highest success.

Guttman analysis. A Guttman scale is a type of scale where items can be ranked in an order, so that individual's response patterns to the scale's items can be ascertained by a single score. More specifically, a Guttman scale is a deterministic scale and as such is based on the assumptions that: (a) all items lower than an individual's ability level were scored correctly, and (b) all items higher than the individual's ability level were scored incorrect. Green (1956) outlined two key statistics required to assess whether a series of tasks fit a scalable pattern. First, the coefficient of reproducibility represents the proportion of original responses which could be reproduced from the single item index, and must be over .90 for items to be considered scalable. Second, the coefficient of consistency indexes the extent to which observed scalability was greater than what is expected by chance alone, and must exceed .50 to be considered significant. The proposed five-item scale failed to meet criteria for a Guttman scale (index of reproducibility = .93, index of consistency = .41). However, a four-item scale (removing the most difficult 'Hidden Emotions' task) met criteria (index of reproducibility = .96, index of consistency = .60). As such, in a younger sample, a four-item rather than five-item scale provides a deterministic measure of ToM development.

Rasch analysis. A Rasch scale is probabilistic and, as such, is based on the assumption that from a person's given ability level, the individual probably responded incorrectly to all higher items, and probably scored correctly for all lower items. Data was analysed using the WINSTEP computer program (Linacre, 2005). Consistent with the methodology of Peterson and colleagues (2005), the item measures were rescaled to give the False Belief task an arbitrary item difficulty measure score of 5.0 on the scale. From this analysis we examined both infit and outfit statistics for both item and person. Item fit statistics assess how well each item fit within the scale. The standardised infit statistic is more sensitive to responses that do not fit the pattern near a person's ability level (e.g., an incorrect score at a lower level than their ability). The outfit statistic is more sensitive to unexpected responses that are further away from the person's measurement level (e.g., an incorrect response to the easiest task and a correct response to the more difficult task). Both statistics have an expected value of 0 and a standard deviation of 1. Fit values of greater than 2.0 indicate the item is a misfit (Wright & Masters, 1982) and thus, does not fit with the proposed scale. A fit value of greater than -2.0 indicates overfit and thus suggests the scale is more deterministic than predicted by the Rasch model (making it a better fit for a Guttman scale). For the overall sample all item fit statistics met criteria, including the mean item fit for the overall scale (see Table 3), indicating the five tasks did indeed form an acceptable Rasch scale. Under a Rasch scale, but not a Guttman scale, one would expect some responses to meet the 'other' category (see Table 2), given the scale is probabilistic, rather than deterministic, in nature.

Similarly, for person fit statistics, a fit value of greater than 2 means the child's response pattern did not adequately fit with a probabilistic model. Only one child's fit statistic showed a pattern of responses that did not fit the statistical model, with a standard infit of 2.0. All other person infit and outfit statistics met criteria for a Rasch model, with an average infit

of -.1 (*SD* = 1.1) and average outfit of 0.0 (*SD* = .7; see Table 3). That is, the response pattern provided by all individuals except one was considered an adequate fit for the Rasch model.

Insert Table 3 here

Rasch analysis also produces item measure statistics that index the difficulty of each item (see Table 3). This analysis showed that the order of observed difficulty matched the rank order of difficulty expected by the scale's authors (Peterson, et al., 2005; Wellman & Liu, 2004). However, the measure statistics of the two most difficult tasks (False Belief and Hidden Emotion) appear comparable in difficulty, suggesting why criteria was not met for the five-item Guttman scale. Specifically, the two tasks differed in difficulty by only 0.12, less than half the standard errors of the estimates (0.28 and 0.29). The close difficulty level of these two tasks was also confirmed by the similar number of children who were able to pass the tasks (see Table 1).

The close measure scores of the two highest items may be due to a floor effect created by the inclusion of the younger sample (i.e., 2 – 3 year olds) who all failed both tasks. Thus, we also performed linear mixed model analyses to test the idea that the difficulty level of the tasks changed with age. We created a logistic mixed-effects model[3] with task outcome (passed or failed) as the dependent variable. Participant ID and item number were added as random effects (random effects ID *SD* = .52; item number *SD* = 1.64). Not surprisingly, the addition of age as a fixed effect significantly improved the fit of the model, $\chi^2(1) = 55.71$, $p = .001$, indicating the likelihood of a child passing the task increased as age increased ($b = .13$, $SE_b = .02$). However, allowing slopes to vary by participant did not improve the fit of the model $\chi^2(2) = 0.76$, $p = .68$. Therefore, there was no evidence of a difference in the difficulty trajectory of the items for the younger or older children in our age range; meaning the similar

[3] All mixed-effects models were created using the lme4 package (Bates, Maechler, & Bolker, 2011) in R, an open-source language and environment for statistical computing (R Development Core Team, 2011).

difficulty levels of the highest two tasks were not due to the inclusion of the younger age range. However, we cannot rule out that the similar levels of task difficulty for the two hardest tasks may be due to our sample on the whole being younger than samples previously used by the scale's authors (e.g., with children up to 12 years old; Peterson, et al., 2012).

In sum, item fit statistics demonstrated good internal consistency between the five tasks, with all tasks fitting the Rasch model. These results, along with evidence of most children being able to pass the control questions on the easier tasks, provide support for the suitability and usability of this scale with this younger age range. While there was evidence of the two highest tasks being equivalent in difficulty, this was not a result of the inclusion of toddlers, but may have been due to the overall younger sample. Regardless, further analysis of the Guttman scale showed a four-item scale (excluding Hidden Emotion) was perhaps more appropriate for use with toddlers.

Convergent Validation

The Sally-Anne false-belief task is often considered the standard assessment of ToM (Bloom & German, 2000). Thus, performance on this task was used to assess the convergent validity of the ToM scale. Convergent validity was assessed by correlating the total number of tasks scored correct[4] with performance on the Sally-Anne task. A strong correlation (r = .67, $p < .001$) showed good convergent validity. As the total number of tasks scored correct increased, so did the likelihood of passing the Sally-Anne task. Thus, the ToM scale provides a score (total items correct) that strongly reflects performance on an already established ToM measure. However, the 45% of shared variance between the two measures clearly demonstrates that the measures are not redundant.

[4] Rasch person measure scores were not used as they correlated with the total items correct at a level of $r = .97$ ($p < .001$).

Discussion

ToM is a key cognitive construct required for typical social development. We argued two key issues limit the ability to draw robust conclusions on ToM development (particularly in the pre-school years). The first was the paucity of empirical evidence on the psychometric properties of ToM measures, and the second was the use in most research of a false-belief task as the single indicator of ToM ability. Our results provide the first demonstration that the five-item ToM scale (Peterson, et al., 2005) is suitable for use with children from two years of age. Further, we found evidence that, while the five-items did form an adequate Rasch scale, a four-item scale is more advisable for use with toddlers.

Evidence of the suitability of this test for use with younger children was confirmed through two key findings. First, from two years of age, children showed comprehension of the scenarios presented in the first three tasks. Second, from two years of age some children showed evidence of diverse-desires and diverse-belief understanding, showing the usefulness of the scale in discriminating between the early ToM abilities of toddlers.

Further to its usability as a single ToM test, we found clear evidence of convergent validity of the five-item scale against the single-index Sally-Anne task. The strength of the correlation between these measures shows that scale score did indeed provide a score that could indicate performance on an alternate, widely used ToM measure. However, the fact the association was not perfect demonstrated the scale also provided a novel measure of early ToM, beyond what can be captured by a false-belief task alone. The earlier (easier) tasks clearly provide a more extensive examination of ToM, prior to the development of false-belief ability.

Practical Applications and Future Direction

To date, the focus on false-belief ability has restricted our assessment, and thus understanding, of early ToM ability. The confirmation of the scale's suitability with toddlers

opens an important window to more comprehensive examination of early ToM development. As the tasks were able to form a scale, they particularly demonstrate the strong potential of the use of this test as a longitudinal ToM measure. Wellman and colleagues (2011) have shown that patterns of responses evidenced in the scales cross-sectional data are replicated longitudinally with late pre-school and school-aged children. Results of the current study now pave the way for the longitudinal investigation of this test in a much younger age range (i.e., from two years old).

Potential limitations of our study were the small and culturally homogeneous sample. As such, future research should focus on the use of this scale with larger samples, particularly of toddlers, as well as children from different cultures (as has been done with slightly older samples; Wellman, Fang, Liu, Zhu, & Liu, 2006). Moreover, it would be useful for future research to explore the scalability and validity of the American version of the five-item scale with this younger age range (Wellman & Liu, 2004). Given ToM's link to social development, the future assessment of the test as a longitudinal assessment would also be highly beneficial in advancing our ability to early detect children who may be at-risk of developing social impairments. If we could identify what specific early ToM abilities (from the five-items) predicted later social difficulties it may open the window for identifying those children who may benefit from early social skills training, or even social support in the pre-school and early-school setting.

Summary

Our research answered some important outstanding questions regarding the use and psychometric properties of the five-item ToM scale. The test was deemed suitable for children from two years of age. While the test did meet requirements for a five-item Rasch scale, our results suggest (depending on the purpose of the test) that a four-item scale would

be more appropriate with two- and three-year old children, particularly given that few were able to comprehend the scenario presented in the most difficult task (Hidden Emotion) and no child under 46 months showed this more complex ToM ability. The good internal consistency and convergent validity of the scale point to the usefulness of viewing the tasks as a single test of ToM, providing researchers access to a validated ToM test, which provides information across six levels of ToM performance.

References

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

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. J. (2000). *Understand others minds: Perspectives from developmental cognitive neuroscience*. New York, US: Oxford University Press.

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. *Computer software.* Retrieved from *http://CRAN. R-project. org/packag e0lme4*.

Bloom, P., & German, T. P. (2000). A brief report: Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*, B25-B31.

Burack, J. A., Charman, T., Yirmiya, N., & Zelazo, P. R. (2001). *The development of autism: Perspectives from theory and research*. New Jersey, US: Lawrence Erlbaum Associates, Inc.

Caputi, M., Lecce, S., Pagnin, A., & Banerjee, R. (2012). Longitudinal effects of theory of mind on later peer relations: The role of prosocial behavior. *Developmental Psychology, 48*(1), 257.

Devine, R. T., & Hughes, C. (2012). Silent films and strange stories: Theory of mind, Gender, and social experiences in middle childhood. *Child Development, 84*(3), 989-1003.

Green, B. F. (1956). A method of scalogram analysis using summary statistics. *Psychometrika, 21*(1), 79-88.

Hughes, C., & Leekam, S. (2004). What are the links between theory of mind and social relations? Review, reflections and new direction for studies of typical and atypical development. *Social Development, 13*(4), 590-619.

Hutchins, T. L., Prelock, P. A., & Bonazinga, L. (2012). Psychometric evaluation of the theory of mind inventory (ToMI): A study of typically developing children and children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 42*(3), 327-341.

Linacre, J. M. (2005). A user's guide to WINSTEPS MINISTEP Rasch-model computer programs. Chicago: Winsteps. com.

Mayes, L. C., Klin, A., Tercyak, K. P., Cicchetti, D. V., & Cohen, D. J. (1996). Test-retest reliability for false-belief tasks. *Journal of Child Psychology and Psychiatry, 37*(3), 313-319.

Onishi, K. H & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 305*(5719), 255-258.

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

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or asperger syndrome. *Child Development, 83*(2), 469-485.

R Development Core Team. (2011). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lyndiche.

Renouf, A., Brendgen, M., Parent, S., Vitaro, F., David Zelazo, P., Boivin, M., . . . Seguin, J. R. (2010). Relations between theory of mind and indirect and physical aggression in kindergarten: Evidence of the moderating role of prosocial behaviors. *Social Development, 19*(3), 535-555.

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology, 47*(5), 1239-1247.

Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science, 17*(12), 1075-1081.

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.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523-541.

Wellmann, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false beliefs. *Child Development, 72*(3), 655-684.

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

Table 1

Number of Children Who Passed the Control Question and Focal Question Along With Age (in Months) at Which Questions Were Passed or Failed

Task	*n* passed (% of *N*)	*M* Pass Age (*SD*)	Youngest Pass	Oldest Pass	Youngest Fail	Oldest Fail
	Control (Preliminary) Question					
Diverse Desire	68 (100)	44.08 (11.26)	24.00	61.00	-	-
Diverse Belief	66 (96)	44.56 (11.08)	24.00	61.00	26.00	31.00
Knowledge Access	44 (65)	49.32 (9.08)	27.00	61.00	24.00	53.00
False-Belief	42 (61)	50.43 (7.89)	32.00	61.00	24.00	53.00
Hidden Emotion	46 (68)	50.28 (7.37)	32.00	61.00	24.00	46.00
Sally-Anne	60 (88)	44.77 (11.36)	24.00	61.00	26.00	55.00
	Focal (Theory-of-mind) Question					
Diverse Desire	53 (78)	47.26 (9.86)	26.00	61.00	24.00	53.00
Diverse Belief	39 (57)	47.85 (9.42)	27.00	60.00	24.00	61.00
Knowledge Access	28 (41)	52.39 (6.36)	33.00	60.00	27.00	61.00
False-Belief	11 (16)	53.09 (4.06)	45.00	60.00	32.00	61.00
Hidden Emotion	9 (13)	55.56 (4.36)	46.00	60.00	32.00	61.00
Sally-Anne	29 (42)	53.17 (6.01)	37.00	61.00	24.00	56.00

Table 2

The Six Scalable Response Patterns and Descriptive Statistics for Responses Fitting Them

Pattern	Diverse Desires	Diverse Beliefs	Knowledge Access	False Belief	Hidden Emotion	%(*n*)
1	-	-	-	-	-	16 (11)
2	+	-	-	-	-	13 (9)
3	+	+	-	-	-	19 (13)
4	+	+	+	-	-	16 (11)
5	+	+	+	+	-	5 (3)
6	+	+	+	+	+	3 (2)
Other						28 (19)

Note. + represents correct response, - represents incorrect response

Table 3

Item and Person Measure Summary and Fit Statistics for Rasch Analysis of Five ToM Tasks

Tasks and Person	Measure	Error	Standardised infit	Standardised outfit
Item difficulty summary				
Hidden Emotion	5.12	.29	.0	-.3
Content False Belief	5.00	.28	.1	-.3
Knowledge Access	3.61	.23	-1.3	-.3
Diverse Beliefs	2.61	.25	1.5	1.3
Diverse Desires	0.86	.38	-.2	-.5
M	3.44	.29	.0	.0
SD	1.59	.05	.9	.7
Person ability summary				
M	2.93	1.14	-.1	.0
SD	1.67	.29	1.1	.7

Note. Expected values for standardised infit and outfit is $M = 0$ and $SD = 1$; a fit statistic > 2.0 indicates a misfit.