


## ARTICLE

# Development of a theory of mind assessment for children using multidimensional Rasch modelling

Shih-Chieh Lee<sup>1,2</sup> | Cheng-Te Chen<sup>3</sup> | I-Ning Fu<sup>1,4</sup> | Meng-Ru Liu<sup>4</sup> |  
Kuan-Lin Chen<sup>5,6,7</sup> 

<sup>1</sup>School of Occupational Therapy, College of Medicine, National Taiwan University, Taipei, Taiwan

<sup>2</sup>Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan

<sup>3</sup>Institute of Learning Sciences and Technologies, National Tsing Hua University, Hsinchu, Taiwan

<sup>4</sup>Child Developmental Assessment & Intervention Center, Zhongxing Branch, Taipei City Hospital, Taipei City, Taiwan

<sup>5</sup>Department of Occupational Therapy, College of Medicine, National Cheng Kung University, Tainan City, Taiwan

<sup>6</sup>Institute of Allied Health Sciences, College of Medicine, National Cheng Kung University, Tainan City, Taiwan

<sup>7</sup>Department of Physical Medicine and Rehabilitation, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan City, Taiwan

## Correspondence

Kuan-Lin Chen, Department of Occupational Therapy, College of Medicine, National Cheng Kung University, Tainan City, Taiwan.  
Email: [klchen@mail.ncku.edu.tw](mailto:klchen@mail.ncku.edu.tw)

## Funding information

National Science and Technology Council (formerly the Ministry of Science and Technology), Taiwan, Grant/Award Number: 110-2628-B-006-031, 109-2628-B-006-023 and 109-2811-B-006-517

## Abstract

Existing theory of mind (ToM) measures for children meet challenges from three perspectives. Developmentally, they lack items covering the entire spectrum of ToM abilities, namely, the early, basic and advanced levels. Dimensionally, most measures are unidimensional, not distinguishing between cognitive and affective ToM. Practically, most measures rely heavily on verbal abilities or lack engaging formats. This study aimed to address these critical issues by developing a Theory of Mind Assessment (ToMA). The items were generated based on classical scenarios spanning all developmental levels. The responses from 574 neurotypical children aged 37 to 194 months were analysed using the multidimensional Rasch model. Ten items showed satisfactory model fits when classified into cognitive (4 items) and affective (6 items) domains, with 16 misfit items excluded. Two items seemed easier for girls and two for boys, and the item difficulties were adjusted. The mean Rasch person reliabilities were 0.78 and 0.81. The scores exhibited small and high correlations with subjective and objective criteria. The newly developed measure may offer reliable, valid and sex-unbiased assessments while satisfying developmental, multidimensional and practical requirements. It seems promising for application in clinical and research settings and is worthy of future refinement and validation to provide high-quality ToM assessment.

## KEYWORDS

multidimensional, psychometric property, Rasch analysis, theory of mind

Theory of mind (ToM) is an ability to infer others' mental states (Premack & Woodruff, 2010), encompassing intentions, desires, beliefs and emotions (Green et al., 2008). Accurately understanding others' mental states aids in the interpretation of social behaviours and guides appropriate responses for

smooth social interactions (Green et al., 2008). Growing evidence indicates that individuals with various diagnoses, such as autism spectrum disorders and schizophrenia, exhibit deficits in ToM, considerably explaining their social dysfunction (Cotter et al., 2018). Previous findings preliminarily suggest that clinical intervention can improve patients' ToM levels and promote their motivation to engage in interactions with others (Dyrda et al., 2020; Holeva et al., 2024; Kurtz & Richardson, 2012). Furthermore, the development of ToM correlates with the development of social function, independent living and overall quality of life (Hughes & Leekam, 2004; Peterson et al., 2016). Accordingly, ToM assessments are a crucial prerequisite for research on and clinical management of individuals with ToM deficits, particularly children.

Remarkable numbers of ToM measures are available for children both with and without diagnoses (Fu et al., 2023). According to another review article (Osterhaus & Bosacki, 2022), the most frequently used ToM measures for children and adolescents include the Strange Stories (Happé, 1994), the Reading the Mind in the Eye Test (Baron-Cohen et al., 2003), the series targeting higher-order false belief (Perner & Wimmer, 1985), the Faux Pas Task (Baron-Cohen et al., 1999) and the Triangles Test (Castelli et al., 2000). However, previous studies have revealed low correlations among these representative ToM measures (Hayward & Homer, 2017; Warnell & Redcay, 2019), implying that these measures may not assess the same ability. Although the findings may also be a consequence of correlation attenuation due to low reliability or inconsistent factor structures across ages (Osterhaus & Bosacki, 2022), these explanations raise concerns about the reliability and validity of these ToM measures. Thus, users are advised to select the most appropriate measure depending on the research questions based on a documented framework (Osterhaus & Bosacki, 2022).

To resolve the limitations, an important prerequisite is to identify the problems. Based on a recent systematic review (Fu et al., 2023), these limitations can be addressed from three perspectives: developmental, multidimensional and practical. From the developmental perspective, few ToM measures include items assessing the entire spectrum of ToM levels (Fu et al., 2023), namely, the early (developed at 1–2 years old), basic (developed for preschool age) and advanced (developed for school age) levels. For instance, the well-known Sally-and-Anne test assesses the understanding that others' beliefs may be counter to reality, also known as first-order false belief, which is developed at the basic level (Baron-Cohen et al., 1985). In contrast, the Faux Pas Task assesses the awareness of embarrassments in social interaction, an ability developed at the advanced level (Stone et al., 2003). Consequently, the ToM scores provided by these measures are incomparable due to the mismatched ToM levels, leading to heterogeneity in research findings. If items assessing different ToM levels can be included in a single measure with jointly calibrated item parameters, such a measure can promote the comparability of findings across studies and thus be suitable for longitudinal investigations in children of different ages. In addition, measures with all representative ToM items can ensure the comprehensiveness of ToM assessments, given the possibility that these ToM measures may reflect distinct abilities that cannot be combined (Warnell & Redcay, 2019). Therefore, there is a need for a ToM measure that includes items assessing the entire spectrum of developmental levels.

From the multidimensional perspective, numerous factor structures have been proposed to support the score calculations and interpretations of ToM assessments. One structure, for example, targets the developmental stages of ToM (Hutchins et al., 2014), which include three factors, namely, the early, basic and advanced domains. Another structure can be constructed depending on the content (Hutchins et al., 2014), such as classifying the same ToM items into emotion recognition, mental state comprehension and pragmatics domains. A third structure is replicable in children across ages (Osterhaus et al., 2016), which captures three mentalizing processes: social reasoning, reasoning about ambiguity and recognizing the transgression of social norms. The last structure focuses on the nature of mental states, distinguishing between cognitive ToM (CToM) and affective ToM (AToM), where CToM involves inferences of mental states irrelevant to emotions and AToM involves emotion-related mental states (Dvash & Shamay-Tsoory, 2014; de la Osa et al., 2016). Among these structures, the distinction between CToM and AToM should be particularly highlighted because they involve different brain regions (Dvash & Shamay-Tsoory, 2014; de la Osa et al., 2016),

indicating that they may represent distinct abilities. However, the empirical evidence for separating CToM and AToM is absent, leaving the validity of this structure unknown. Thus, at a minimum, optimized ToM measures should separate CToM and AToM into different domains to improve the construct validity.

From the practical perspective, many of these measures are inaccessible to children with language impairments in comprehension and expression (Fu et al., 2023). For one thing, most measures present scenarios using spoken stories, sometimes with accompanying photos, which can be challenging for those with difficulty in language comprehension (Fu et al., 2023). For another, the reliance on open-ended questions in many measures poses challenges to children with difficulty in language expression (Fu et al., 2023). Accordingly, the ToM scores may underestimate the actual abilities of these children, as they may possess the knowledge but struggle to articulate it due to their limited language abilities. Furthermore, the validity of these ToM scores may be constrained due to significant correlations with language functions (Osterhaus & Bosacki, 2022). To address this issue, ToM measures should minimize the influences from language function and optimize the utility in children across different conditions.

The identified limitations can be resolved through the following methods to improve the validity and appropriateness of assessing children's ToM abilities. Regarding the developmental perspective, including items that span the entire spectrum of ToM development in a measure is crucial. As for the multidimensional perspective, the CToM and AToM items should be divided into different domains. It is also important to examine the underlying structure to ensure the unidimensionality of each domain (Lee, Lin, Huang, et al., 2021; Lee et al., 2023; Wellman & Liu, 2004). Considering the practical perspective, presenting the items with a series of images could offer more accessible information and might be a compromise solution for two reasons. First, measures with visual prompts may decrease interference from difficulties in language comprehension (Garcia-Molina & Clemente-Estevan, 2019; Sperotto, 2016). Second, videos may have negative effects on children's performance on ToM tasks (Anderson & Pempek, 2005; Reiß et al., 2017).

Furthermore, the Rasch model helps optimize the measurement properties of ToM measures for three reasons. First, the Rasch model yields interval scores, enabling the interpretation of differences in magnitude (Hays et al., 2000). Second, it provides an individual reliability index, facilitating the interpretation of the findings (Hays et al., 2000). Third, it allows the developers to examine the inconsistent item difficulties for boys and girls, thereby ensuring fair assessment (Lee et al., 2022; Zieky, 2003). In addition, the Rasch model is also available for multidimensional constructs, making it well suited for constructing a ToM measure. In conclusion, the multidimensional Rasch model is an ideal solution to optimize ToM measures, as it satisfies all considerations from developmental, multidimensional and practical perspectives.

This study aimed to develop a ToM assessment (ToMA) to satisfy the requirements from the three aforementioned perspectives simultaneously. Regarding the developmental requirement, the ToMA should include items tailored to all levels of ToM development. To meet the dimensional requirement, the ToMA should employ a multidimensional model to delineate CToM and AToM. To satisfy the practical requirement, each ToMA item should be presented using a series of images, and the responses should be obtained through forced-choice questions. To further optimize measurement properties, the ToMA was developed based on the multidimensional Rasch model, particularly ensuring the fairness of assessments between boys and girls. Therefore, the ToMA could be useful for monitoring and comparing ToM developments across preschool and school-age children.

## METHOD

### Transparency and openness

This study's design and its analysis were not preregistered.

## Participants

This prospective study analysed data obtained from three ongoing research projects. Two projects aimed to develop and validate the ToMA items designed for preschoolers and school-age children with eight and 13 items (Fu, 2017; Fu et al., 2024; Liu, 2020), respectively. The remaining project aimed to develop the ToMA items for the entire spectrum of ToM development with a wider age band. We collected the data through different projects because responding to all the ToMA items would be burdensome and could lead to cognitive fatigue in children, as they might face a series of difficult items that they were unable to answer due to incomplete development in these abilities. To minimize the negative influences on children, all the ToMA items were administered to children in only one of the projects. The participants were recruited from preschools, elementary schools and after-school care facilities in Tainan City, a major population center in southern Taiwan. The inclusion and exclusion criteria were the same across the three projects; that is, children were included if they were regarded as neurotypical by both parents and teachers. The only exception was that the children's ages differed across projects, with children aged 3–6 (Fu, 2017; Fu et al., 2024), 6–12 (Liu, 2020) and 3–12 years old targeted in the respective projects. Children within this age band were targeted because most ToM abilities develop in this period. Regardless of the projects, the participants were excluded if they had a diagnosis of any neurodevelopmental disease (e.g. autism spectrum disorder), had uncorrectable visual or auditory impairment, or were unfamiliar with Mandarin Chinese. The data collection was approved by the Institutional Review Board of the National Cheng Kung University Hospital (BR-105-020-T).

A total of 574 children, with roughly equal numbers (50.3% vs. 49.7%) of boys and girls, participated in this study. Approximately 39.5% and 34.7% of the children completed the sets of ToMA items for preschoolers (8-item version) and school-age children (13-item version), while 25.8% of the children completed the items for the entire spectrum of development. The mean age was 82.2 months, with ages ranging from 37 to 194 months. They were at the early and basic levels of development of ToM competence and performance. Details are listed in Table 1.

## Procedures

The children and their caregivers were invited to participate in this study and received an explanation of the procedures and purposes. Oral assent and informed consent were obtained from the children and their caregivers if they agreed to participate. Then the children were assessed with the Verbal Comprehension Index (VIC), the ToMA items and the external ToM measure (i.e. Theory of Mind Task Battery), while their caregivers completed a questionnaire on demographic characteristics and the Theory of Mind Inventory–Second Edition. All assessments were completed in a quiet room and administered by licensed occupational therapists familiar with these measures. The ToMA was administered using a computer program, and the children responded to the questions by pressing the option buttons.

## Measures

To satisfy the requirement from the developmental perspective, the ToMA items were designed based on the most classical and representative scenarios used in the existing ToM measures for children aged 3–12 (Fu et al., 2023). The selected scenarios included emotion distinction (2–3 years) (Sivaratnam et al., 2012), diverse desires (3–4 years) (Wellman & Liu, 2004), first-order false belief (including unexpected content and location design, 4–6 years) (Baron-Cohen et al., 1985), second-order false belief (6–8 years) (Perner & Wimmer, 1985), lie (7–8 years) (Happe, 1994), white lie (7–8 years) (Broomfield et al., 2010; Happe, 1994), irony (7–8 years) (Filippova & Astington, 2008; Happe, 1994), third-order

TABLE 1 Characteristics of the children ( $n = 574$ ).

Characteristic	Value
Boys, $n$ (%)	289 (50.3)
Age (months), Mean (SD)	82.2 (29.7)
VCI, Mean (SD)	110.7 (14.2)
ToMTB, Mean (SD)	
Advanced	2.6 (1.9)
Basic	2.8 (1.5)
Early	4.5 (0.7)
Total	9.9 (3.6)
ToMI-2, Mean (SD)	
Advanced	12.9 (4.4)
Basic	16.6 (2.3)
Early	17.3 (1.7)
Total	15.6 (2.4)

Abbreviations: ToMI-2, Theory of Mind Inventory–Second Edition; ToMTB, Theory of Mind Task Battery; VCI, Verbal Comprehension Index.

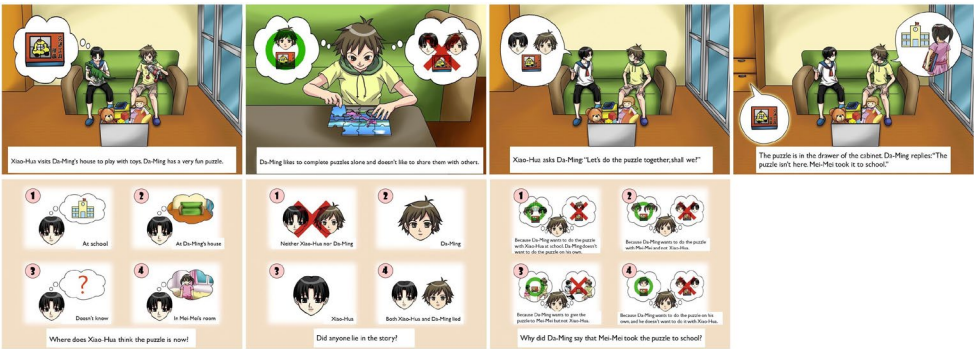


FIGURE 1 An example of an image series for a CToM item (C\_L1I).

false belief (8–10 years) (Nettle & Liddle, 2006), real apparent emotion (8–10 years) (Happe, 1994) and faux pas (9–11 years) (Baron-Cohen et al., 1999).

Regarding the dimensional requirement, the characteristics and stories for CToM and AToM items were different to avoid biases caused by assessing CToM and AToM using the same story. The differences between CToM and AToM with the same scenario can be seen in Figures 1 and 2, which present the unexpected content items as examples. The domains of the ToMA items were determined by the inferred mental states related to CToM and AToM, in accordance with evidence reported in previous studies (Dvash & Shamay-Tsoory, 2014; de la Osa et al., 2016).

As for the practical requirement, the ToMA items were scored based on three forced-choice questions: justification (whether the children demonstrate the targeted ToM abilities), control (whether the children understand the scenario), and confirmation (whether the children's responses are based on ToM abilities and not just guesses) questions. The responses to these questions were used to determine the item score within a binary scale, wherein the points would be awarded if all questions were answered correctly. In addition, a description and cartoon images were composed for each item to maximize the understandability and accessibility to the targeted children.





FIGURE 2 An example of an image series for an AToM item (A\_LI1).

This study analysed a total of 26 items classified into two domains: CToM (7 items) and AToM (19 items). The content validity and suitability of the items for children aged 3–12 years were reviewed and approved by an expert panel. In addition, the understandability and accessibility of the ToMA items for the targeted children were confirmed using the cognitive debriefing method.

The VCI was used to evaluate the children's language function. Depending on whether the children were younger or older than 6 years, the index was obtained from the Taiwanese versions of the Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition (consisting of the Information and Similarities subtests) (Wechsler, 2012), or the Wechsler Intelligence Scale for Children—Fourth Edition (comprising the Vocabulary, Similarities and Comprehension subtests) (Wechsler, 2003). Both measures demonstrate good test–retest reliability (coefficients ranging from 0.72 to 0.94), reliability and validity in neurotypical children (Chen & Chen, 2007, 2013).

The Theory of Mind Task Battery and the Theory of Mind Inventory—Second Edition were used as external criteria regarding the concurrent validity (Hutchins et al., 2014). These two measures were used because they reflect different ToM concepts (ToM competence in a standardized environment vs. ToM performance in real-life contexts). The Theory of Mind Task Battery comprises 15 items embedded in 9 vignettes. The items are rated dichotomously (as fail or pass), and the total score range is 0–15 points. The Theory of Mind Inventory—Second Edition, on the other hand, contains 60 caregiver-reported items. The items are rated using a 20-point visual analogue scale with labels, 'definitely not (zero points)' and 'definitely (20 points)', at the opposite ends. Both measures provide a total score and three domain scores representing the early, basic and advanced levels of ToM. We used the Taiwanese versions of the Theory of Mind Task Battery and the Theory of Mind Inventory—Second Edition, as these versions have shown adequate reliability (Cronbach's  $\alpha > 0.90$  and intraclass correlation coefficient for test–retest reliability  $> 0.86$ ) and validity (Pearson's correlation coefficient between both measures, 0.48) (Chen et al., 2023; Chiu et al., 2016). In the current study, the Cronbach's  $\alpha$  values of the Theory of Mind Task Battery and the Theory of Mind Inventory—Second Edition were 0.84 and 0.81, respectively.

## Data analysis

We initially examined the data–model fits of the multidimensional Rasch model using both infit and outfit mean squares (Wright & Linacre, 1994). The expected values of both statistics are 1.0. A value larger than 1.0 generally indicates that children's responses do not fit well to the Rasch model (i.e. misfit), while a value smaller than 1.0 represents that children's responses are predictable (i.e. overfit) (Wright & Linacre, 1994). We defined misfit items as those showing any fit statistic value exceeding 1.2 (Wright & Linacre, 1994), and such items were removed iteratively until the remaining items showed satisfactory model fits. The correct percentages were further considered to ensure

the validity of all retained items. However, the overfit items, which exhibited any fit indices smaller than 0.8, were retained to include as many of the ToMA items as possible. The item coverage was demonstrated using an item–person map. The Rasch analysis was conducted in ConQuest—Second Edition (Wu et al., 2007).

Then we examined the differential item functioning (DIF) of sex for the remaining items by comparing the item difficulties of the ToMA items estimated for boys and girls, separately (Zieky, 2003). The existence of DIF of sex was determined by DIF values and Z score significance (Lee et al., 2022). DIF values exceeding a range of  $\pm 0.38$  indicated a large and unneglectable difference (Zieky, 2003). Z score, a ratio of the DIF values to the corresponding SE, exceeding a range of  $\pm 1.96$  was regarded as a significant difference because the differences could not be explained by random error of the estimations (Lee et al., 2022). The item difficulties were adjusted according to the children's sex if both the DIF value and Z score indicated the presence of a DIF of sex (Lee et al., 2022).

Principal component analysis was conducted to examine whether any common factors underlay the standardized residuals of the Rasch analysis (Linacre, 1998). The existence of a common factor was determined by eigenvalues. An eigenvalue larger than three was regarded as a cut-off because it is the lowest requirement to form a common factor (Costello & Osborne, 2005).

The Rasch person reliability was calculated for each child. The mean Rasch person reliability was also used to assess the overall performance of the ToMA for group-level comparisons. The percentages of children obtaining Rasch person reliability over 0.90 and 0.70 were calculated to describe the performance of the ToMA in individual-level comparisons (Lee, Lin, Liu, et al., 2021), where reliability coefficients over 0.90 and 0.70 were regarded as good and acceptable (Aaronson et al., 2002), respectively. Note that the imputation was adopted because the data were combined across three research projects wherein the numbers of ToMA items administered to the children were inconsistent and would have been biased with the raw data.

Pearson's correlation coefficient (Pearson's  $r$ ) was used to examine the concurrent validity and even the inter-domain correlations. Pearson's  $r$  values over 0.1, 0.4 and 0.7 were considered to indicate low, moderate and high correlations (Akoglu, 2018), respectively. High correlations were anticipated for the ToMA and the total score of the Theory of Mind Task Battery compared to the Theory of Mind Inventory—Second Edition because both the ToMA and the Theory of Mind Task Battery assess ToM competence. In contrast, low correlations were expected between ToMA and VCI scores, as they could be considered evidence of divergent validity.

## RESULTS

The 26 ToMA items were analysed. After the removal of 16 items, the remaining items showed satisfactory model fits (infit statistics = 0.88–1.15; outfit statistics = 0.58–1.11). The percentages of correct responses are presented in Table A1. Four AToM items demonstrated substantial DIF of sex (DIF values =  $-2.67$  to  $-2.35$  and  $1.50$  to  $1.82$ ). Among them, two DIF items (i.e. emotion distinction and diverse desire) were easier for girls than for boys and two items (i.e. lie and white lie) were easier for boys, as shown in Table 2. Residual-based principal component analysis showed that no common factors underlay the standardized residuals of the ToMA items, with the largest eigenvalue being 1.8.

The distributions of the item difficulties generally covered the targeted children, except the distribution of children with low and moderate levels (CToM =  $-6.4$  to  $-1.2$  logits; AToM =  $-4.4$  to  $0.4$  logits) of ability (Figure 3). The average Rasch reliabilities of the CToM and AToM were 0.78 and 0.81, respectively. Approximately 90% of the children obtained individual Rasch reliabilities over 0.70 for the CToM and AToM domains, while about 55% achieved individual Rasch reliabilities over 0.80.

Correlational analysis showed extremely high inter-domain correlations between CToM and AToM scores ( $r = 0.99$ ). The ToMA scores demonstrated high correlations with the Theory of Mind Task Battery (both  $r$ s of the CToM and AToM were 0.80 when rounded to two decimal places), while those with the

TABLE 2 Model fit, DIF of sex and adjusted item difficulty for boys and girls in the ToMA items.

Domain	No.	Item	Outfit statistics	Infit statistics	DIF value	Z score	Item difficulty (boy)	Item difficulty (girl)
CToM	1	C_TF1	1.04	1.15	0.27	1.36	1.72	
	2	C_SF1	0.74	0.92	0.03	0.32	0.43	
	3	C_LI1	1.01	1.02	−0.08	0.70	−0.81	
	4	C_UC1	0.77	0.90	−0.23	1.71	−1.34	
AToM	1	A_WL3	0.59	0.88	1.74	3.84	2.92	
	2	A_UL1	1.10	1.10	−0.04	0.24	1.39	
	3	A_WL1	0.85	0.94	1.50	7.98	2.68	−0.32
	4	A_LI1	1.02	0.98	1.82	9.73	2.37	−1.27
	5	A_DD1	1.11	1.06	−2.35	11.85	−3.59	1.10
	6	A_ED1	0.58	0.88	−2.67	10.56	−7.46	−2.11

Abbreviations: AToM, Affective theory of mind; CToM, Cognitive theory of mind; DD, Diverse Desires; DIF, Differential Item Functioning; ED, Emotion Distinction; IR, Irony; LI, Lie; RA, Real Apparent; SF, Second-order False Belief; TF, Third-order False Belief; UC, Unexpected Content; UL, Unexpected Location; WL, White Lie.

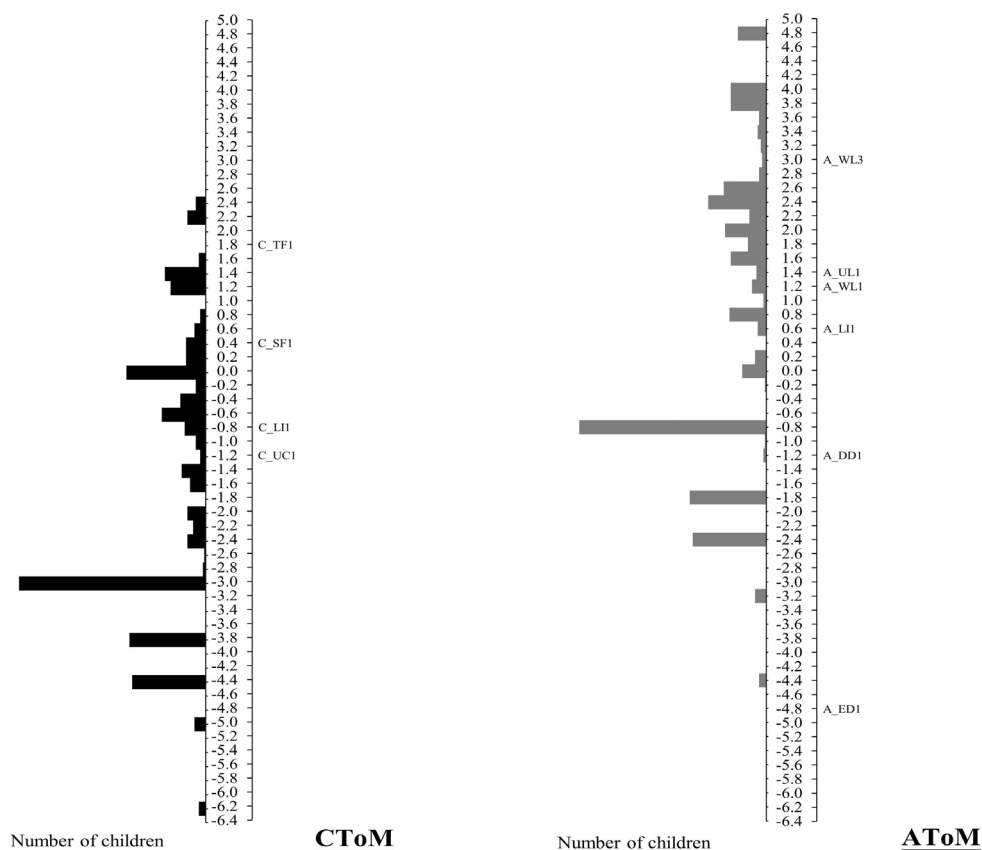
Theory of Mind Inventory—Second Edition were moderate (both *r*s of the CToM and AToM were the same, 0.37). Low correlations (*r*s = 0.15 and 0.16) were found between the ToMA and VCI scores.

## DISCUSSION

This study aimed to develop a reliable and valid ToM measure based on the requirements from the developmental, multidimensional and practical perspectives. The findings show that the newly developed ToMA may provide generally reliable, valid and feasible assessments of CToM and AToM abilities, although the separation of CToM and AToM abilities was not strongly supported by the empirical data. The ToMA shows great applicability for both clinical and research users, for it includes the most classical and representative scenarios covering the entire spectrum of ToM development levels (from early to advanced); it offers interval scaling and an individualized reliability index due to its development based on an advanced testing model (the Rasch model); and it presents items using cartoon images with forced-choice questions to minimize the barriers caused by language impairment. Accordingly, the ToMA seems to be a promising measure assessing CToM and AToM in children across developmental stages, and it is worthy of future refinement and validation to optimize its psychometric properties. Moreover, its development may shed some light on the complexity of ToM structures for children across developmental stages, as the current study offers empirical evidence for separating CToM and AToM into different domains.

Ten ToMA items showed satisfactory model fits to the multidimensional Rasch model. In addition, the largest eigenvalue provided by the residual-based principal component analysis was small, suggesting that no common factors existed across the residuals of ToMA items. The findings support that the ToMA items are suitable for the Rasch model because the children's responses to the items matched the Rasch model's assumptions, including the unidimensionality of each domain. Moreover, the findings also indicate that the parameters estimated based on the Rasch model are valid, including item difficulties and two ToM scores. Furthermore, the high correlations with the Theory of Mind Task Battery and the moderate correlations with the Theory of Mind Inventory—Second Edition support the convergent validity of the ToMA. The findings also imply that the ToMA scores may be more likely to reflect explicit ToM rather than applied ToM. In contrast, the low correlation with the VCI suggests that the ToMA is not dominated by language function and supports its divergent validity. Based on these findings, the validity of the CToM and AToM scores is preliminarily supported for children across developmental stages.





**FIGURE 3** The item-person map of the CTOM and ATOM domains. ATOM, Affective theory of mind; CTOM, Cognitive theory of mind; DD, Diverse Desires; ED, Emotion Distinction; IR, Irony; LI, Lie; RA, Real Apparent; SF, Second-order False Belief; TF, Third-order False Belief; UC, Unexpected Content; UL, Unexpected Location; WL, White Lie.

The extremely high inter-domain correlations challenge the claim that CTOM and ATOM are distinct abilities. However, several factors may affect the correlations and could be considered before treating the ToMA items as unidimensional. First, most studies suggest that ToM is likely to be multidimensional rather than unidimensional (Hutchins et al., 2014; Osterhaus et al., 2016; Osterhaus & Koerber, 2021; Warnell & Redcay, 2019), although diverse structures have been proposed across studies. In addition, fitting to a unidimensional Rasch model resulted in more misfit items and substantially lower Rasch person reliability, which is consistent with a previous study and implies that the structure is inadequate (Osterhaus et al., 2016). Second, the correlations might have been overestimated for several reasons: (1) the ToMA points were awarded if examinees correctly answered three forced-choice questions, (2) the children's ToM development was ongoing and thus related to cognitive factors (Gallant et al., 2020) and (3) the scores were estimated using a multidimensional Rasch model, which takes inter-domain correlations into account (Wang et al., 2004; Wang & Chen, 2004). Third, high inter-domain correlations do not always indicate the same domains; for example, competence and performance are highly correlated but distinct constructs (Lee et al., 2014, 2024; Wade et al., 2018). Considering these factors, CTOM and ATOM may be theoretically distinguishable but empirically mixed. To avoid misleading users, we tentatively selected the two-factor structure for the ToMA.

The average Rasch reliabilities of the two ToM abilities were acceptable to modest. Moreover, about 90% of the children achieved individual reliabilities over 0.70. These results indicate that the ToMA is

sufficient for group-level comparisons (Aaronson et al., 2002), which are often conducted in research settings, such as those for intervention and control groups. However, the ToMA's reliability was lower than the requirement for individual-level comparisons (Aaronson et al., 2002), such as those in clinical settings targeting the ToM abilities of individual children. Therefore, conservative interpretations of the ToMA scores are warranted for clinical practitioners to avoid misguided clinical decisions. Fortunately, the Rasch model contributes an individual reliability index for each ToM score. If the individual reliability of a score is low, the ToMA can be retested to obtain a more reliable result to improve the efficacy of the ToM assessment.

The modest reliability for individual-level comparisons may be attributed to two factors. First, the lack of items suitable for children with low to medium levels of CToM and AToM may have led to limited information about their abilities, leading to lower reliability. Second, our participants were neurotypical children who rarely have difficulty in the development of ToM abilities. Thus, the reliability of the ToMA items might have been reduced because of the constrained variations in ToM abilities (Wang & Chen, 2004). Future studies may include more items to improve reliability. For example, joint attention items can be considered to belong to the CToM domain because they are conceptually easier than unexpected context items (Fu et al., 2023). Regarding the AToM domain, items that are more difficult than emotion distinction but easier than diverse desire items are warranted. Moreover, future studies may also include children with diagnoses, such as children with autism spectrum disorders, because their inclusion would facilitate exploration of the variations of ToM abilities and increase the reliability of the ToMA. After such exploration, the ToMA could be more precise and feasible for children with diverse ToM abilities.

Four items showed substantial DIF of sex, demonstrating that these ToMA items are not equally difficult for boys and girls. The findings imply that comparing ToM scores between boys and girls may be unfair and even biased if DIF items are not recognized without controls for their influences. Our findings may partially explain the female advantage in social-related assessments reported in previous studies (Greenberg et al., 2023; Kirkland et al., 2013), highlighting the importance of examining and controlling the DIF of sex on ToM assessments, as these examinations are rarely mentioned in previous studies. A possible contributor may be related to the social experiences in daily interactions (Lee et al., 2022). For instance, girls may have more opportunities to express their emotions and be more likely to care about others' feelings. Thus, girls may be more familiar with these kinds of tasks, which might have led to the lower item difficulty. In contrast, boys have higher frequencies of telling lies compared to girls (Gervais et al., 2000; Guerra et al., 2022; Lee, 2013), which may partially explain their familiarity with these tasks. Although the actual causes of the DIF items remain unclear, our findings suggest that DIF of sex may be more relevant to emotion-related tasks, as all the DIF items were AToM and not CToM items. Nevertheless, we controlled the influences of DIF of sex on the ToMA items by adjusting the item difficulties in accordance with the examinees' sex (Lee, Lin, Liu, et al., 2021). Accordingly, the ToMA can provide sex-unbiased assessments and improve the fairness of the ToM assessments for boys and girls.

Three advantages and their corresponding implications of the ToMA can be noted. First, the ToMA provides interval scores of children's ToM abilities, which can substantially improve the quality of ToM assessments, wherein most ToM items are rated dichotomously (Lee et al., 2024). Therefore, the ToMA scores can demonstrate the magnitude of differences in individuals directly, allowing clinical practitioners to compare the ToM levels between two children or the scores obtained from the same child within repeated assessments. This advantage also makes the ToMA useful for researchers because interval scores are a prerequisite for most parametric statistical analyses. Second, the ToMA provides individual Rasch reliabilities and SEs for each score. Thus, users can feasibly determine whether the results of current assessments are reliable and consider retesting if the reliability is unsatisfactory. Moreover, the individual SE is useful for calculating 95% confidence intervals and interpreting whether children significantly improve after receiving interventions. Third, the scenarios of the ToMA items are presented using serial images, which facilitate children's understanding of the items and improve their motivation to complete the assessments (Fu et al., 2023). Therefore, the challenges of administering

ToM assessments can be minimized, and the efficacy of ToM assessments is improved. Based on these advantages, the ToMA is deemed to be a promising measure with implications for both clinical and research settings.

## CONSTRAINTS ON GENERALITY

Some limitations may be considered while interpreting the current findings. First, we included only neurotypical children who had relatively good ToM ability. Thus, the variations in the ToM spectrum and the Rasch person reliability of the ToMA might have been underestimated because no responses from individuals with poor ToM, such as children with autism, were included. Second, we adopted a stringent criterion (MNSQ <1.2) to evaluate the Rasch model fit (Aryadoust et al., 2020). Thus, some classical ToM items might have been deleted, and their item properties might not have been considered in the current study. Third, the Rasch person reliability was estimated based on the imputed data because only about 30% of the children completed the full sets of ToMA items. Accordingly, the generalizability of the findings may have been constrained and may require further cross-validation. Fourth, the ToMA assesses ToM competence only, so it cannot directly reflect children's real-life performances. To resolve this limitation, adding items to evaluate children's ToM performances can be considered, or the ToMA may be jointly analysed with other measures assessing ToM performance, such as the Theory of Mind Inventory–Second Edition.

## CONCLUSION

The newly developed ToMA may provide generally reliable, valid and sex-unbiased assessments of CToM and AToM abilities in neurotypical children based on a tentative two-factor structure. Because the ToMA can satisfy the developmental, multidimensional and practical requirements that hamper the application of the existing ToM measures, it shows great potential for application in both clinical and research settings, and it is worthy of future refinement and validation to improve the quality of ToM assessment.

## STATEMENTS AND DECLARATIONS

The manuscript has been read and approved by all the authors. The manuscript has not been published, nor is it currently under consideration for publication elsewhere.

## AUTHOR CONTRIBUTIONS

**Shih-Chieh Lee:** Conceptualization; methodology; formal analysis; investigation; writing – original draft; writing – review and editing. **Cheng-Te Chen:** Conceptualization; methodology; writing – review and editing. **I-Ning Fu:** Conceptualization; writing – review and editing. **Meng-Ru Liu:** Conceptualization; writing – review and editing. **Kuan-Lin Chen:** Conceptualization; methodology; formal analysis; investigation; writing – review and editing; funding acquisition; resources; supervision.

## ACKNOWLEDGEMENTS

The authors report there are no competing interests to declare. All authors agree with the stated authorship and contributions of this article. This work was supported by the Taiwan National Science and Technology Council (formerly the Ministry of Science and Technology under Grant numbers 110-2628-B-006-031, 109-2628-B-006-023 and 109-2811-B-006-517). The data collection was approved by the Institutional Review Board of the National Cheng Kung University Hospital (BR-105-020-T).

## CONFLICT OF INTEREST STATEMENT

None of the authors have any conflicts of interest to declare.

## DATA AVAILABILITY STATEMENT

Data available on request from the authors.

## ORCID

Kuan-Lin Chen  <https://orcid.org/0000-0001-9996-814X>

## REFERENCES

- Aaronson, N., Alonso, J., Burnam, A., Lohr, K. N., Patrick, D. L., Perrin, E., & Stein, R. E. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research*, 11, 193–205. <https://doi.org/10.1023/a:1015291021312>
- Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18, 91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Anderson, D. R., & Pempek, T. A. (2005). Television and very young children. *American Behavioral Scientist*, 48, 505–522. <https://doi.org/10.1177/0002764204271506>
- Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, 38, 6–40. <https://doi.org/10.1177/0265532220927487>
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition*, 21, 37–46. [https://doi.org/10.1016/0010-0277\(85\)90022-8](https://doi.org/10.1016/0010-0277(85)90022-8)
- Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, 29, 407–418. <https://doi.org/10.1023/a:1023035012436>
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2003). The “Reading the mind in the eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42, 241–251. <https://doi.org/10.1111/1469-7610.00715>
- Broomfield, K. A., Robinson, E. J., & Robinson, W. P. (2010). Children's understanding about white lies. *British Journal of Developmental Psychology*, 20, 47–65. <https://doi.org/10.1348/026151002166316>
- Castelli, F., Happe, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, 12, 314–325. <https://doi.org/10.1006/nimg.2000.0612>
- Chen, K. L., Jiang, D. R., Yu, Y. T., & Lee, Y. C. (2023). Development and psychometric evidence of the Chinese version of the theory of mind Inventory-2 (ToMI-2) in children with autism spectrum disorder. *Research in Autism Spectrum Disorders*, 103, 102132.
- Chen, Y. H., & Chen, H. Y. (2007). *Wechsler intelligence scale for children: Chinese version manual*. Chinese Behavioral Science Corporation.
- Chen, Y. H., & Chen, H. Y. (2013). *Wechsler preschool and primary scale of intelligence—Chinese version manual*. Chinese Behavioral Science Corporation.
- Chiu, W. T., Lee, Y. C., Lin, C. H., Jiang, D. R., Chen, C. T., & Chen, K. L. (2016). A comparison of theory of mind capacity and performance in children with autism spectrum disorder and typical development. *Journal of Taiwan Occupational Therapy Association*, 34, 198–213. [https://doi.org/10.6594/JTOTA.2016.34\(2\).03](https://doi.org/10.6594/JTOTA.2016.34(2).03)
- Costello, A. B., & Osborne, J. (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. <https://doi.org/10.7275/yjy1-4868>
- Cotter, J., Granger, K., Backx, R., Hobbs, M., Looi, C. Y., & Barnett, J. H. (2018). Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neuroscience and Biobehavioral Reviews*, 84, 92–99. <https://doi.org/10.1016/j.neubiorev.2017.11.014>
- Dvash, J., & Shamay-Tsoory, S. G. (2014). Theory of mind and empathy as multidimensional constructs. *Topics in Language Disorders*, 34, 282–295. <https://doi.org/10.1097/tld.0000000000000040>
- Dyrda, K., Lucci, K. D., Bieniek-Pocielej, R., & Brynska, A. (2020). Therapeutic programs aimed at developing the theory of mind in patients with autism spectrum disorders - available methods and their effectiveness. *Psychiatria Polska*, 54, 591–602. <https://doi.org/10.12740/PP/108493>
- Filippova, E., & Astington, J. W. (2008). Further development in social reasoning revealed in discourse irony understanding. *Child Development*, 79, 126–138. <https://doi.org/10.1111/j.1467-8624.2007.01115.x>
- Fu, I. N. (2017). *Development, validation, and applicability of the newly developed Brief Preschool Theory of Mind Assessment (BPToMA)* [Master's thesis, National Cheng Kung University, Tainan, Taiwan].
- Fu, I. N., Chen, C. T., Chen, K. L., Liu, M. R., & Hsieh, C. L. (2024). Development and validation of the newly developed preschool theory of mind assessment (ToMA-P). *Frontiers in Psychology*, 15, 1274204. <https://doi.org/10.3389/fpsyg.2024.1274204>
- Fu, I. N., Chen, K. L., Liu, M. R., Jiang, D. R., Hsieh, C. L., & Lee, S. C. (2023). A systematic review of measures of theory of mind for children. *Developmental Review*, 67, 101061. <https://doi.org/10.1016/j.dr.2022.101061>

- Gallant, C. M. M., Lavis, L., & Mahy, C. E. V. (2020). Developing an understanding of others' emotional states: Relations among affective theory of mind and empathy measures in early childhood. *British Journal of Developmental Psychology*, 38, 151–166. <https://doi.org/10.1111/bjdp.12322>
- Garcia-Molina, I., & Clemente-Estevan, R. A. (2019). Autism and faux pas. Influences of presentation modality and working memory. *Spanish Journal of Psychology*, 22, E13. <https://doi.org/10.1017/sjp.2019.13>
- Gervais, J., Tremblay, R. E., & Desmarais-Gervais, L. (2000). Children's persistent lying, gender differences, and disruptive behaviours: A longitudinal perspective. *International Journal of Behavioral Development*, 24, 213–221. <https://doi.org/10.1080/016502500383340>
- Green, M. F., Penn, D. L., Bental, R., Carpenter, W. T., Gaebel, W., Gur, R. C., Kring, A. M., Park, S., Silverstein, S. M., & Heinssen, R. (2008). Social cognition in schizophrenia: An NIMH workshop on definitions, assessment, and research opportunities. *Schizophrenia Bulletin*, 34, 1211–1220. <https://doi.org/10.1093/schbul/sbm145>
- Greenberg, D. M., Warrier, V., Abu-Akel, A., Allison, C., Gajos, K. Z., Reinecke, K., Rentfrow, P. J., Radecki, M. A., & Baron-Cohen, S. (2023). Sex and age differences in “theory of mind” across 57 countries using the English version of the “Reading the mind in the eyes” test. *Proceedings of the National Academy of Sciences of the United States of America*, 120, e2022385119. <https://doi.org/10.1073/pnas.2022385119>
- Guerra, A., Randon, E., & Scorcu, A. E. (2022). Gender and deception: Evidence from survey data among adolescent gamblers. *Kyklos*, 75, 618–645. <https://doi.org/10.1111/kykl.12305>
- Happe, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24, 129–154. <https://doi.org/10.1007/BF02172093>
- Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, 38, II28–II42. <https://doi.org/10.1097/00005650-200009002-00007>
- Hayward, E. O., & Homer, B. D. (2017). Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. *British Journal of Developmental Psychology*, 35, 454–462. <https://doi.org/10.1111/bjdp.12186>
- Holeva, V., Nikopoulou, V. A., Lytridis, C., Bazinas, C., Kechayas, P., Sidiropoulos, G., Papadopoulou, M., Kerasidou, M. D., Karatsioras, C., Geronikola, N., Papakostas, G. A., Kaburlasos, V. G., & Evangeliou, A. (2024). Effectiveness of a robot-assisted psychological intervention for children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 54, 577–593. <https://doi.org/10.1007/s10803-022-05796-5>
- Hughes, C., & Leekam, S. (2004). What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development. *Social Development*, 13, 590–619. <https://doi.org/10.1111/j.1467-9507.2004.00285.x>
- Hutchins, T. L., Prelock, P. A., & Bouyea, L. B. (2014). *Theory of Mind Inventory & Theory of Mind Task Battery*. <http://www.theoryofmindinventory.com/task-battery>, <http://www.theoryofmindinventory.com/task-battery>
- Kirkland, R. A., Peterson, E., Baker, C. A., Miller, S., & Pulos, S. (2013). Meta-analysis reveals adult female superiority in “Reading the mind in the eyes test”. *North American Journal of Psychology*, 15, 121–146. <https://psycnet.apa.org/record/2013-09240-009>
- Kurtz, M. M., & Richardson, C. L. (2012). Social cognitive training for schizophrenia: A meta-analytic investigation of controlled research. *Schizophrenia Bulletin*, 38, 1092–1104. <https://doi.org/10.1093/schbul/sbr036>
- Lee, K. (2013). Little liars: Development of verbal deception in children. *Child Development Perspectives*, 7, 91–96. <https://doi.org/10.1111/cdep.12023>
- Lee, S. C., Chen, K. W., Huang, C. Y., Li, P. C., Hsieh, T. L., Lee, Y. C., & Hsueh, I. P. (2022). Development of a Rasch-calibrated test for assessing implied meaning in patients with schizophrenia. *The American Journal of Occupational Therapy: Official Publication of the American Occupational Therapy Association*, 76(4), 7604205020. <https://doi.org/10.5014/ajot.2022.047316>
- Lee, S. C., Fu, I. N., Liu, M. R., Yu, T. Y., & Chen, K. L. (2023). Factorial validity of the theory of mind Inventory-2 in typically developing children. *Journal of Autism and Developmental Disorders*, 53, 310–318. <https://doi.org/10.1007/s10803-022-05426-0>
- Lee, S. C., Huang, C. Y., Fu, I. N., & Chen, K. L. (2024). Interpreting the results of explicit and applied theory of mind collectively in autistic children: A solution from Rasch analysis. *Autism*, 28, 355–366. <https://doi.org/10.1177/13623613231170698>
- Lee, S. C., Lin, G. H., Huang, Y. J., Huang, S. L., Chou, C. Y., Chiang, H. Y., & Hsieh, C. L. (2021). Cross-validation of the factorial validity of the stroke impact scale 3.0 in patients with stroke. *The American Journal of Occupational Therapy: Official Publication of the American Occupational Therapy Association*, 75(2), 7502205070p1–7502205070p10. <https://doi.org/10.5014/ajot.2021.040659>
- Lee, S. C., Lin, G. H., Liu, C. C., Chiu, E. C., & Hsieh, C. L. (2021). Development of the CAT-FER: A computerized adaptive test of facial emotion recognition for adults with schizophrenia. *American Journal of Occupational Therapy*, 75, 7501205140p1–7501205140p11. <https://doi.org/10.5014/ajot.2020.043463>
- Lee, Y. C., Chen, S. S., Koh, C. L., Hsueh, I. P., Yao, K. P., & Hsieh, C. L. (2014). Development of two Barthel index-based supplementary scales for patients with stroke. *PLoS One*, 9, e110494. <https://doi.org/10.1371/journal.pone.0110494>
- Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions*, 12, 636. <https://www.rasch.org/rmt/rmt122m.htm>



- Liu, M. R. (2020). *Development, validation and applicability of newly developed Brief School-aged Theory of Mind Assessment (BSc-ToMA)* [Master's thesis, National Cheng Kung University, Tainan, Taiwan].
- Nettle, D., & Liddle, B. (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology*, 4, 231–244. <https://doi.org/10.1556/jcep.4.2006.3-4.3>
- de la Osa, N., Granero, R., Domenech, J. M., Shamay-Tsoory, S., & Ezpeleta, L. (2016). Cognitive and affective components of theory of mind in preschoolers with oppositional defiance disorder: Clinical evidence. *Psychiatry Research*, 241, 128–134. <https://doi.org/10.1016/j.psychres.2016.04.082>
- Osterhaus, C., & Bosacki, S. L. (2022). Looking for the lighthouse: A systematic review of advanced theory-of-mind tests beyond preschool. *Developmental Review*, 64, 101021. <https://doi.org/10.1016/j.dr.2022.101021>
- Osterhaus, C., & Koerber, S. (2021). Social cognition during and after kindergarten: The relations between first-order and advanced theories of mind. *European Journal of Developmental Psychology*, 18, 573–592. <https://doi.org/10.1080/17405629.2020.1820861>
- Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of advanced theory-of-mind tasks. *Child Development*, 87, 1971–1991. <https://doi.org/10.1111/cdev.12566>
- Perner, J., & Wimmer, H. (1985). “John thinks that Mary thinks that...” attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39, 437–471. [https://doi.org/10.1016/0022-0965\(85\)90051-7](https://doi.org/10.1016/0022-0965(85)90051-7)
- Peterson, C., Slaughter, V., Moore, C., & Wellman, H. M. (2016). Peer social skills and theory of mind in children with autism, deafness, or typical development. *Developmental Psychology*, 52, 46–57. <https://doi.org/10.1037/a0039833>
- Premack, D., & Woodruff, G. (2010). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1, 515–526. <https://doi.org/10.1017/s0140525x00076512>
- Reiß, M., Krüger, M., & Krist, H. (2017). Theory of mind and the video deficit effect: Video presentation impairs children's encoding and understanding of false belief. *Media Psychology*, 22, 23–38. <https://doi.org/10.1080/15213269.2017.1412321>
- Sivaratnam, C. S., Cornish, K., Gray, K. M., Howlin, P., & Rinehart, N. J. (2012). Brief report: Assessment of the social-emotional profile in children with autism spectrum disorders using a novel comic strip task. *Journal of Autism and Developmental Disorders*, 42, 2505–2512. <https://doi.org/10.1007/s10803-012-1498-8>
- Sperotto, L. (2016). The visual support for adults with moderate learning and communication disabilities: How visual aids support learning. *International Journal of Disability, Development and Education*, 63, 260–263. <https://doi.org/10.1080/1034912x.2016.1153256>
- Stone, V. E., Baron-Cohen, S., Calder, A., Keane, J., & Young, A. (2003). Acquired theory of mind impairments in individuals with bilateral amygdala lesions. *Neuropsychologia*, 41, 209–220. [https://doi.org/10.1016/s0028-3932\(02\)00151-3](https://doi.org/10.1016/s0028-3932(02)00151-3)
- Wade, M., Prime, H., Jenkins, J. M., Yeates, K. O., Williams, T., & Lee, K. (2018). On the relation between theory of mind and executive functioning: A developmental cognitive neuroscience perspective. *Psychonomic Bulletin & Review*, 25, 2119–2140. <https://doi.org/10.3758/s13423-018-1459-0>
- Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 28, 295–316. <https://doi.org/10.1177/0146621604265938>
- Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9, 116–136. <https://doi.org/10.1037/1082-989X.9.1.116>
- Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, 191, 103997. <https://doi.org/10.1016/j.cognition.2019.06.009>
- Wechsler, D. (2003). *Wechsler intelligence scale for children—fourth edition (WISC-IV)*. The Psychological Corporation.
- Wechsler, D. (2012). *Wechsler preschool and primary scale of intelligence—Fourth edition (WPPSI-IV)*. The Psychological Corporation.
- Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75, 523–541. <https://doi.org/10.1111/j.1467-8624.2004.00691.x>
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370. <https://www.rasch.org/rmt/rmt383b.htm>
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest 2.0: General item response modelling software* [Computer program manual] Australian Council for Educational Research.
- Zieky, M. (2003). *A DIF primer Educational Testing Service*.

**How to cite this article:** Lee, S.-C., Chen, C.-T., Fu, I.-N., Liu, M.-R., & Chen, K.-L. (2025). Development of a theory of mind assessment for children using multidimensional Rasch modelling. *British Journal of Psychology*, 00, 1–15. <https://doi.org/10.1111/bjop.12785>

APPENDIX

TABLE A1 The percentages of correct responses in children of different age groups.

Domain	Item	3y	4y	5y	6y	7y	8y	9y	10y	11y	12y
CToM	C_TF1	0	0	0	0	5	15	32	21	29	37
	C_SF1	0	1	10	15	23	35	44	62	53	56
	C_LI1	6	3	16	33	40	53	68	69	84	59
	C_UC1	6	3	18	33	48	68	85	81	92	85
AToM	A_WL2	0	1	2	8	13	26	32	40	53	41
	A_UL1	18	6	15	38	45	68	68	74	82	81
	A_WL1	6	3	16	32	48	50	76	81	89	89
	A_LI1	6	3	16	33	53	76	90	83	92	89
	A_DD1	24	32	77	80	81	94	100	98	100	100
	A_ED1	76	92	98	98	98	100	100	100	100	100

*Note:* The values presented in this table are percentages of correct responses in each age group, which were calculated based on the data inputted with expected values from the multidimensional Rasch model to make the findings comparable.

Abbreviations: AToM, affective theory of mind; CToM, cognitive theory of mind; DD, Diverse Desires; ED, Emotion Distinction; IR, Irony; LI, Lie; RA, Real Apparent; SF, Second-order False Belief; TF, Third-order False Belief; UC, Unexpected Content; UL, Unexpected Location; WL, White Lie.