Si Wang, Glenda Andrews, Donna Pendergast, David Neumann, Yulu Chen & David H. K. Shum (2022) A Cross-Cultural Study of Theory of Mind Using Strange Stories in School-Aged Children from Australia and Mainland China, Journal of Cognition and Development, 23:1, 40-63.

A Cross-Cultural Study of Theory of Mind Using Strange Stories in School-Aged Children from Australia and Mainland China

To date, cross-cultural studies on Theory of Mind (ToM) have predominantly focused on pre-schoolers. This study focuses on middle childhood, comparing two samples of mainland Chinese (n = 126) and Australian (n = 83) children aged between 5.5 and 12 years. Strange Stories, the most commonly used measure of ToM, was employed. The study aimed to: examine the one-factor versus two-factor structure and measurement invariance of Strange Stories across two cultures; use the verified invariant model of Strange Stories to compare children's cognitive and affective ToM across two cultures; and finally, to investigate correlates of individual differences on Strange Stories cross-culturally. Multiple-groups confirmatory factor analysis revealed the measurement invariance of a two-factor model of Strange Stories (cognitive and affective) in both groups. The results revealed that mainland Chinese children had an equal performance on cognitive ToM stories and poorer performance on affective ToM stories compared to their Australian counterparts. Cultural differences in the factors related to individual differences of ToM. The number of older siblings was a positive predictor of mainland Chinese school-aged children's cognitive ToM, and contrarily was a negative predictor in the Australian sample. The findings confirm that Strange Stories is a reliable measure for evaluating ToM in school-aged children from mainland China and Australia and highlights the importance of considering the cognitive and affective aspects of ToM in cross-cultural comparison.

Keywords: theory of mind, culture, strange stories, middle childhood

Theory of Mind (ToM) refers to the ability to attribute mental states, including knowledge, beliefs, and intentions to oneself and to others (Premack & Woodruff, 1978). It develops quickly in the first few years of childhood. Understanding of false beliefs is a well-accepted marker of children's ToM in early childhood (Andrews et al., 2003; Wellman et al., 2001). Unlike false-belief understanding, which focuses on the cognitive component of ToM, other measures for school-aged children incorporate both cognitive and affective components, both of which develop further during middle childhood (Cassetta et al., 2018; Wilson et al., 2018). Compared with research in preschool children, there is a gap in research

on ToM development in middle childhood (Peterson, 2004), especially studies from a cross-cultural perspective. Cross-cultural comparisons are necessary because social-cultural factors, such as cultural values and parenting practice, play important roles in children's ToM (Shahaeian et al., 2014; Wang et al., 2016). In the following review of literature, an examination of the ways in which Strange Minds has been applied to research in middle childhood is examined. This includes an examination of the difference between cognitive ToM and affective ToM, following which studies undertaken with a focus on cultural comparison of ToM are considered. Finally, findings from studies with an interest in the influence of siblings are examined. This sets a clear scene for the design of this present study.

Assessing Theory of Mind using Strange Stories in Middle Childhood

The most commonly used measure of ToM for school-aged children is Strange Stories (Happe, 1994) which assesses one's ability to explain the characters' behavior or reaction in specific social scenarios (Devine & Hughes, 2013, 2016; White et al., 2009). The original Strange Stories measure contains 12 types of stories: Lie, White Lie, Joke, Pretend, Misunderstanding, Persuade, Appearance/Reality, Figure of Speech, Sarcasm, Forget, Double Bluff, and Contrary Emotions. At the end of each story, participants are asked two questions. The first takes the form, 'Was it true what X said?' and the second takes the form, 'Why did X say that?'. The participant's answer to the second question is scored as either referring to mental states or physical states.

Previous research assumed that Strange Stories assesses children's ToM as a unitary construct and found children's performance on Strange Stories developed with age (Lecce et al., 2019; O'Hare et al., 2009) and was highly related with their verbal ability (Devine & Hughes, 2016; Lecce et al., 2017). O'Hare et al. (2009) assessed 5- to 12-year-old children's performance on all 12 types of Strange Stories. They found that Strange Stories discriminated ToM development through middle childhood. To reduce the variances of difficulty in stories,

some researchers administered a subset of Strange Stories in which each story was not too difficult but sufficiently challenging. For example, Fletcher et al. (1995) used a subset of Strange Stories in a positron emission tomography study in adults and found that the medial frontal gyrus on the left showed a specific pattern of activation when comparing participants' brain activities on mental state stories with physical stories. This subset consisted of White Lie, Persuasion, Double Bluff, and Misunderstanding, with two instances of each type. White et al. (2009) used the same subset of stories and reported that 7- to 11-year-old children with autism performed more poorly than typically developing children. Devine and Hughes (2013) used one Misunderstanding, one Double Bluff, and two White Lie stories to measure ToM in 8- to 13-year-olds. After controlling for verbal ability, ToM was shown to increase significantly with age. Next, Devine and Hughes (2016) employed two Misunderstanding, two Double Bluff, and one Persuasion stories in a sample of 460 children aged 7 to 13 years. They confirmed that participants' scores correlated with both age and verbal ability. Recently, based on a longitudinal study which followed children from 9.5 to 10.5 years, Lecce et al. (2017) reported children's performance on a short version Strange Stories (two Double Bluff, one Persuasion, one Misunderstanding, and one White Lie) improved across time and significantly correlated with verbal ability. Later, Lecce et al. (2019) reported 9-year-olds performed significantly worse than children aged 10 to 12 years on a subset of Strange Stories that comprised of two Double Bluff, two Misunderstanding, and one Persuasion.

Both the whole set and subset of Strange Stories are confirmed to be a reliable measure of individual differences in ToM across middle childhood. However, the assumption that all the stories load on a unitary construct may be too simplistic. According to Wilson et al. (2018), these stories vary in affective tone according to specific social scenarios (Wilson et al., 2018). These researchers divided the stories into two categories. Joke, Pretend,

Misunderstanding, Appearance Reality, Figure of Speech, Forget, and Double Bluff stories were categorized as low affective tone stories. Meanwhile, Lie, White Lie, Sarcasm, Persuasion, and Contrary Emotion stories, were categorized as high affective tone stories. They found that 5- to 12-year-old Australian children's scores on high and low affective tone stories were differentially associated with cognitive functions. By using regression models including age, verbal ability, cool and hot EF components, verbal ability, working memory, and delay aversion predicted performance on low affective tone stories. In contrast, age and gift delay were significant predictors in performance on high affective tone stories. These results are compatible with a possible differentiation between cognitive and affective ToM within the Strange Stories (Shamay-Tsoory & Aharon-Peretz, 2007).

Cognitive ToM versus Affective ToM

According to Shamay-Tsoory et al. (2010), cognitive ToM focuses on what a person is thinking, whereas affective ToM focuses on how a person is feeling. Cognitive ToM is a prerequisite for affective ToM, and the integration of cognitive ToM and empathy enables a functioning affective ToM. The two constructs have been shown to have different neural correlates. While cognitive ToM is related to activity in the dorsolateral prefrontal cortex, affective ToM is associated with activity in the ventromedial prefrontal cortex (Kalbe et al., 2010; Shamay-Tsoory et al., 2006). A recent neuroimaging meta-analysis study showed that there are three groups of neurocognitive processes underpinning both ToM and empathy, which are predominantly cognitive processes, more affective processes, and combined processes which engage cognitive and affective functions in parallel (Schurz et al., 2020). Furthermore, evidence from gender differences in ToM among school-aged children reveals the general earlier development in cognitive ToM than affective ToM. Lonigro et al. (2014) suggested that girls aged from 9 to 10 years showed earlier development in affective ToM, but not cognitive ToM, than boys in the same age group. Similarly, Cassetta et al. (2018)

found that affective second-order false belief understanding was more advanced in 8- to 11-year-old girls than boys, while there was no gender difference in the cognitive task. As a result, to understand ToM comprehensively, it is important to distinguish between cognitive and affective ToM.

Cultural Comparison of Theory of Mind in Middle Childhood

Comparing ToM in children from different cultures is essential to inform the issue of the extent to which ToM is an innate and culturally universal construct as opposed to a culturally-specific phenomenon. There are many reasons why ToM might differ between Western and Eastern cultures, especially for the Chinese culture. For example, mainland China's collectivistic culture and Western individualistic culture have different influences on the construal of self and others (Triandis, 2001). People in individualistic cultures see themselves as independent, autonomous, and distinct units and they tend to assert their individuality. People in collectivistic cultures value interpersonal relationships and tend to build up their self-identity and personal interests based on their group membership (Bochner, 1994; Markus & Kitayama, 1991). As a result, Chinese children are educated to honor elders, cherish knowledge, and pay more attention to external factors (e.g., rules, guidance) and self-control. In contrast, children raised in most Western countries learn more about personal autonomy and expression of individual opinions and internal mental states (Slaughter & Perez-Zapata, 2014).

The findings from previous studies suggest the presence of cultural differences in ToM in early childhood. For example, while both children from North America and China pass the false belief task around 4 to 5 years old (Liu et al., 2008), children from Hong Kong (Liu et al., 2008) and Japan (Naito & Koyama, 2006) showed a 2 year lag on false-belief understanding. Cultural differences also emerged in preschoolers' sequence of acquisition of various aspects of ToM. Chinese children pass the Knowledge Ignorance task before the

Diverse Beliefs task (Wellman et al., 2006). Comparatively, American and Australian children pass the Diverse Belief task before the Knowledge Ignorance task (Shahaeian et al., 2011; Wellman et al., 2011). This reflects that Eastern cultures value pragmatic knowledge while Western cultures value personal ideas and beliefs.

Fewer studies have examined cultural differences in ToM for school-aged children, although there is some evidence that cultural differences persist into middle childhood. Kobayashi et al. (2007) examined the neural correlates of second-order false belief understanding among 8- to 12-year-old Japanese-English bilingual and English-speaking monolingual children. Although there was no difference in the behavioral results, the fMRI study showed that the monolingual group had stronger activations in the right inferior parietal lobe and the overlapping temporoparietal junction, whereas the bilingual group had stronger activations in the left superior temporal sulcus and the overlapping temporal pole. The results suggest that the neural correlates of ToM differ depending on children's cultural/linguistic background. Furthermore, Shahaeian et al. (2014) recruited a group of 7- to 9-year-old Iranian and Australian children and found that Iranian children showed a better understanding of sarcasm than their Australian peers. They explained that Iranian collectivist culture's emphasis on sacrificing individual needs for group purpose may facilitate children's sensitivity to the hidden meaning involved in sarcasm.

A recent study directly compared ToM in school-aged children from the United Kingdom and Hong Kong (Wang et al., 2016). The results showed that compared to British children, children attending local schools in Hong Kong, performed poorer on ToM tasks. This is consistent with the delayed false belief understanding in Hong Kong preschoolers. However, Hong Kong may not be representative of Eastern cultures. For instance, Hong Kong is more westernized and impacted by British culture (Bond & Cheung, 1983), and Hong Kong children are more likely to be bilingual (Tardif & Chan, 2005). Hong Kong

parents perceived themselves as less warm and more controlling than parents in mainland China (Berndt et al., 1993). Therefore, a question remains regarding whether there is any difference of ToM between mainland Chinese and Western school-aged children.

The Role of Siblings in Theory of Mind

Evidence from Western studies has shown that pre-schoolers' false belief understanding is positively related to different aspects of the sibling structure, such as the total number of siblings (Perner et al., 1994), numbers of older siblings (Ruffman et al., 1998), or siblings aged from 1 to 12 years (McAlister & Peterson, 2007; Peterson, 2000). Pre-schoolers with more siblings, especially older siblings, may outperform "only-children" on false-belief task because the former have more exposure to family-based experiences from siblings and parents, which, in turn, benefits the understanding of others' mental states (Prime et al., 2016; Ruffman et al., 1998). However, fewer studies have focused on how siblings might impact school-aged children's ToM and no consistent results have been revealed. Miller (2013) reported no significant relationship between sibling composition and 5- to 8-year-olds' second-order false-belief understanding. Lecce et al. (2017) reported that school-aged children's performance on the Strange Stories or Silent Film task was not correlated with the number of siblings. In contrast, Kennedy et al. (2015) found that among children aged from 4 to 11 years, those with more older siblings or same-sex siblings demonstrated better performance on the interpretive ToM task. Compared with Western children, mainland Chinese children are likely to be the only child or the older sibling in the family because of the one-child policy (from 1979 to 2015) and the two-child policy (started from 2016) in China. Thus, it is reasonable to assume that kin-relationships and friendships should be more relevant to Chinese children' ToM rather than the sibship. Only one previous study found that the number of older cousins and the frequency of playing with cousins was negatively related to Chinese preschoolers' false belief understanding (Lewis et al., 2006).

The Present Study

The Strange Stories task is a widely-used ToM measure in middle childhood. This present study used a subset of six stories as the measure of school-aged children's ToM. Misunderstanding, Double Bluff, Appearance Reality are categorized as cognitive ToM stories because all scenarios require little interpretation of the character's emotion. White Lie, Sarcasm, and Persuasion are categorized as Affective ToM stories because they require participants to firstly understand the feeling of the characters and then interpret the intention of the utterance. Before examining cultural differences in performance on Strange Stories between mainland Chinese and Australian children, it is necessary to verify the measurement invariance across cultures to establish that there are equivalent relations with no systematic biases between each story and the latent factor(s). If the assumption is invalid, cross-cultural comparisons can be misleading (Hughes et al., 2014). Multiple-groups confirmatory factor analysis (MGCFA) is an important and useful statistical method to test measurement invariance across groups (Fischer & Karl, 2019) and was the method adopted in this study.

Therefore, the first aim of the present study was to examine the one-factor versus two-factor structure and measurement invariance of Strange Stories across two cultures by using MGCFA. The subset of six stories was administrated to 5.5-to 12-year-old children from mainland China and Australia. We hypothesized that a two-factor model (cognitive ToM and affective ToM) would produce a better fit than an overall one-factor model. Meanwhile, the subset of Strange Stories would meet the requirement of measurement invariance and be suitable for assessing ToM across mainland Chinese and Australian samples.

The second aim of the present study was to use the verified invariant model of Strange Stories to compare children's cognitive and affective ToM across two cultures. Given that the timing of passing false belief understanding task is equivalent in preschoolers from mainland China and North America (Liu et al., 2008), we hypothesized that both mainland Chinese and

Australian children would show parallel age-related differences in the performance on cognitive ToM stories. In contrast, because Chinese culture emphasizes self-control and emotion constraint to accommodate the social group to which the individual belongs (Markus & Kitayama, 1991; Matsumoto et al., 2008), Chinese children may have a lower frequency of expression of personal emotions compared with children from Western cultures. As a result, we hypothesized that Chinese children would perform poorer than their Australian counterparts on affective ToM stories.

The last aim of the study was to investigate correlates of individual differences on Strange Stories in both cultural samples. Strange Stories have been found to correlate significantly with age and verbal ability (Devine & Hughes, 2016). Gender differences have been found in affective ToM rather than cognitive ToM (Cassetta et al., 2018). Furthermore, having siblings, especially older siblings, is predictive of children's false belief understanding in early childhood (Prime et al., 2016). Based on these previous results, we hypothesized that for children in both cultural groups, age and verbal ability would predict both cognitive and affective ToM. Gender would only predict children's affective ToM, while the number of older siblings would only predict children's cognitive ToM.

Method

Participants

Participants were recruited through one primary state school in Brisbane (Queensland), which has a population of 2.51 million in 2019 (Australian Bureau of Statistics, 2020), and one public school in a city in the central area of China (Xinyang, Henan) with a population of 6.47 million in 2018 (Henan Province Bureau of Statistics, 2019). The schools were selected because both are public schools located in middle-size cities in their countries. Information sheets and consent forms of the study were sent home with students through their schools. Parents were asked to provide consent and additional information about each child to the

researchers prior to the study to ensure they met the eligibility criteria. Also, parents were asked to complete a questionnaire to collect background information about the child's family. Participants were treated under the National Statement of Ethical Conduct in Research involving Humans, and institutional ethics approval was obtained.

There were 209 participants aged from 5.5 to 12 years, consisting of 83 Australian children (43 males, M_{age} = 8.65, SD = 1.82) and 126 mainland Chinese children (74 males, M_{age} = 8.75, SD = 1.65). There was no significant difference in age between Australian and mainland Chinese samples ($p > .05$). All mainland Chinese children were born in China and spoke Mandarin. All Australian participants were born in Australia and spoke English. In the Australian sample, all parents identified as being Australian with 81.9% being Caucasian and born in Australia (n = 68), 12.0% born in mainland China (n = 10), and 6.0% born in India (n = 5). Parents' education levels were reported and was coded as 1= year 6 and less, 2 = year 10, 3 = year 12, 4 = some university, 5 = bachelor, and 6 = postgraduate. Maternal and paternal education level were summed for each participant to create a parental education index potentially ranged from 2 to 12. The parental education index was used as an indicator of the SES of the child's family.

Procedure

The scripts and coding guidelines of Strange Stories were prepared in English first, and the researcher adopted a back-translation approach (Brislin, 1970). Each word was translated into Mandarin and then back-translated into English by a group of three Chinese/English bilingual developmental psychology researchers (one PhD candidate, two PhDs of psychology). The group discussed and revised the Chinese version to ensure that the Chinese version was equivalent to the original English version. Participants completed the subset of Strange Stories task and verbal test in the following fixed order: White Lie,

Sarcasm, Persuasion, Misunderstanding, Double Bluff, Appearance Reality, and Peabody Picture Vocabulary Test.

Measures

Demographic Questionnaire

A parent-report questionnaire was designed to gather demographic information about the children and their family, including children's gender, age, ethnicity, parents' education levels, and their occupations, and the number of older siblings, younger siblings, playmates and family size which reflected the number of people living at the participant's home.

Theory of Mind Task: Strange Stories

A subset of Strange Stories from O'Hare et al. (2009) was used to measure school-aged children's ToM in this study. The stories were White Lie, Sarcasm, Persuasion, Misunderstanding, Double Bluff, and Appearance Reality. A detailed description of stories is provided in the supplementary materials. Each story presented a character who, within the context of a particular scenario, says something untrue. Participants were firstly asked a comprehension question, "Is it true, what …. said?" to check basic understanding of the story. A second justification question, "Why did she/he say this?" assessed ToM. Children's justifications were coded based on the degree to which the individual responds in terms of the character's psychological state. Responses were scored from 0 to 2, with 0 indicating an incorrect or physical state response, 1 indicating a partial psychological state response, and 2 indicating a full and accurate psychological state response based on O'Hare et al. (2009)'s published guidelines. The score for each story ranged from 0 to 2 and was treated as ordinal variables. The first author scored the whole sample and another coder randomly selected 10% of participants' scripts in both cultural groups and independently scored all the responses. The kappa coefficients for the 6 stories (in the order listed above) were as follows: $\kappa = 1.00$,

κ = .85, κ = 1.00, κ = 1.00, κ = 1.00, and κ =.92. Disagreements were resolved through discussion.

Verbal Ability.

Peabody Picture Vocabulary Test-Fourth Edition (PPVT-4). The PPVT-4 (Dunn et al., 2007) was used to measure receptive vocabulary in the Australian sample. An array of four colored pictures was presented for each vocabulary item. The experimenter asked the child to point to the picture in the array that matched the spoken word. The child's response was scored as correct (1) or incorrect (0). The items were arranged in sets of 10 items that were intended to become increasingly difficult. A basal-set was established when a set contains one or no errors within one set of 10 items. The child continued until the ceiling-set was reached which means a set of 10 items containing eight or more errors. A child's raw score is the number of correct answers below the ceiling set, ranges from 0 to 228. Split-half reliability ranged from .90 to .97 for test ages 5-11 years, based on normative data on monolingual English-speaking children (Dunn et al., 2007).

Peabody Picture Vocabulary Test-Revised Edition (PPVT-R). The Chinese version of PPVT-R (Sang & Miao, 1990) was used for mainland Chinese children. This version is an adaptation of the PPVT-R (Dunn & Dunn, 1981). The representation of the picture and children's response mode are the same as those for PPVT-4. There were 175 items in total. The first item was determined by the child's PPVT age. If the child made 8 or more correct responses before the first error, a basal was established. A ceiling was reached when a child incorrectly identified six of eight consecutive items. The ceiling was defined as the last item in the lowest series of eight successive items with six incorrect responses. A child's raw score is the number of correct answers below the ceiling item, ranges from 0 to 175. The Chinese version of PPVT-R's test-retest reliability was .94, and its split-half reliability was .98 (Sang & Miao, 1990).

Results

Data Treatment

A multi-group confirmatory factor analysis (MGCFA) was conducted to investigate the latent factor structure of Strange Stories and its measurement invariance. The analyses were conducted using the "lavaan" package in R (Version 4.2). The MGCFA is effective in construct validation and the assessment of measurement invariance (Fischer & Karl, 2019). Given that the score of each story could be treated as an ordinal variable, the mean- and variance-adjusted weighted least squares estimator (WLSMV) was used in the MGCFA models (Li, 2016). The adequacy of model fit is judged by the following criteria: Chi-square test of model fit is not significant; root mean square error of approximation (RMSEA) is less than .06; both comparative fit index (CFI) and Tucker–Lewis index (TLI) exceed .90 (Hu & Bentler, 1998). For the model comparison, a significant chi-square difference indicates that model fit is significantly worse in the more restricted model (Anderson & Gerbing, 1988), and the differences in fit indices of RMSEA, CFI, and TLI are equal to or less than .01 (Little et al., 2007). The effect sizes (Cohen's *d*) were interpreted according to recommendations from Kline (2005): small standardized effect sizes ranged from .10 to .30, moderate effect sizes ranged from .30 to .50, and large effect sizes were greater than .50.

Descriptive Analysis of the Two Samples

The descriptive statistics for age, older and younger siblings, playmates, family size, verbal ability, parental education level, and score on each story are shown in Table 1. Culture had a significant relationship with the number of older siblings, $\chi^2(3) = 28.54$, $p < .001$, Cramer's $V = .37$, and younger siblings, $\chi^2(3) = 16.43$, $p = .001$, Cramer's $V = .28$. Australian children had more siblings than mainland Chinese children. There was no difference of the number of playmates, $t(206) = 0.20$, $p = .84$, and family size, $t(204) = 0.58$, $p = .56$, between the two cultural samples. Besides, parental education level of Australian

sample was higher than that of mainland Chinese sample, $t(205) = 4.22$, $p < .001$. The raw scores for the stories were significantly intercorrelated in the whole sample, except there was no significant correlation between Appearance Reality and Sarcasm (see Table S1 in the supplementary materials). Although some of the correlations were not significant in mainland Chinese and Australian sample, the expected directions were aligned with those in the whole sample (see Table 2).

Table 1. Descriptive statistics of age, siblings, verbal ability, parental education level, and scores of each story across cultures

	mainland China			Australia		
	n	*M(SD)*	*Range*	*n*	*M(SD)*	*Range*
Age (years)						
Male	74	8.87 (1.64)	5.50 - 11.58	43	8.34 (1.87)	5.58 - 11.92
Female	52	8.58 (1.66)	5.83 - 11.50	40	8.99 (1.72)	5.58 - 11.83
Total	126	8.75 (1.65)	5.50 - 11.58	83	8.65 (1.82)	5.58 - 11.92
Siblings						
Older	126	.21 (.44)	0 - 2	83	.66 (.72)	0 - 3
Younger	126	.42 (.51)	0 - 2	83	.73 (.75)	0 - 3
Playmates	125	4.68 (2.04)	0 - 10	83	4.59 (4.40)	0 - 25
Family size	123	4.31 (1.35)	2 - 11	83	4.41 (0.99)	2 - 7
Verbal ability	126	135.96 (19.18)	88 -163	83	153.52 (23.77)	93-210
Parent EDU	126	8.31 (1.82)	2 -12	81	9.54 (2.37)	3 -11
Cognitive ToM stories						
MU	126	1.45 (0.69)	0 - 2	83	1.42 (0.91)	0 - 2
Double Bluff	126	1.19 (0.76)	0 - 2	83	1.27 (0.73)	0 - 2
AR	126	1.14 (0.72)	0 - 2	83	1.35 (0.85)	0 - 2
Affective ToM stories						
White Lie	126	1.69 (0.54)	0 - 2	83	1.86 (0.47)	0 - 2
Sarcasm	126	0.73 (0.80)	0 - 2	83	0.90 (0.86)	0 - 2
Persuasion	126	1.45 (0.82)	0 - 2	83	1.41 (0.87)	0 - 2

Note. Verbal ability was represented by children's PPVT raw scores. Parental EDU = Parental education level, MU = Misunderstanding, AR = Appearance Reality.

Table 2. Correlations among key study variables in Australian (Above diagonal) and mainland Chinese (Below diagonal) samples

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Age	-	.79***	.18	-.01	-.25*	.30**	.21+	.01	.43***	.36***	.21+	.32**	.43***	.51***
2. Verbal ability	.79***	-	-.01	.16	-.29**	.21+	.16	-.09	.34**	.43***	.24*	.38***	.45***	.46***
3. Gender	-.09	-.10	-	-.05	-.05	.12	.01	.11	.14	.01	.09	-.01	-.06	.10
4. Parental EDU	-.03	.05	-.07	-	.03	.10	-.01	.20+	.09	-.10	.21+	.05	.11	.02
5. Older sibling	-.04	.02	.01	-.38***	-	-.46***	.00	.40***	-.19+	-.38***	-.20+	-.32**	-.15	-.22*
6. Younger sibling	.02	.02	.13	.03	-.28**	-	-.05	.46***	.13	.15	.09	.10	.13	.09
7. Playmates	.08	.13	-.02	.05	.03	-.13	-	-.01	.05	-.01	.10	.05	.26*	.10
8. Family size	-.10	-.08	.08	-.14	-.06	.50***	.09	-	-.03	-.12	-.03	-.21+	-.05	-.13
9. MU	.44***	.44***	.06	-.13	.16+	-.11	.11	.04	-	.23*	.30**	.26*	.30**	.21+
10. Double Bluff	.21*	.30***	.07	.07	-.12	-.11	.12	.02	.40***	-	.24*	.29**	.33**	.27*
11. AR	.20*	.29**	.04	-.07	.23**	-.10	.00	-.04	.34***	.15+	-	.37***	.13	.20+
12. White Lie	.35***	.39***	.27**	-.02	-.03	-.05	.08	-.03	.29***	.13	.28**	-	.20+	.35**
13. Sarcasm	.50***	.43***	-.06	.05	.02	-.09	.24**	.01	.35***	.19*	.04	.16+	-	.35**
14. Persuasion	.48***	.50***	.01	-.08	.03	-.08	.04	-.11	.33***	.17+	.16+	.28**	.33***	-

Note. $^{+}p < .10$; $*p < .05$; $** p < .01$; $*** p < .001$. Gender was coded as: 0 = male, 1= female. Parental EDU = Parental education level, MU = Misunderstanding, AR = Appearance Reality

One-factor versus Two-factor Model of Strange Stories across Cultures

First, the one-factor model of Strange Stories was tested in the whole sample. The model showed an unacceptable model fit, $\chi^2(9) = 19.16$, $p = .02$, RMSEA = 0.07, CFI = 0.92, TLI = 0.87. The model fit was improved by permitting residuals between Appearance Reality and White Lie (standardized estimate = .203, $p = .011$) to be correlated. In Appearance Reality story, Mr. Brown told Alice he was Santa instead of Mr. Brown. Most children who provided partial and correct psychological explanation assumed that Mr. Brown said this because he wanted to make Alice believe Santa exists and not to hurt her feeling. This is very similar to with the prosocial mentalizing concept of White Lie (Cheung et al., 2015), in which a person told a lie to protect others' feeling. This provided a reasonable justification for the modification of permitting residuals between the two stories. Therefore, after adjusting the model by adding this residual, the one-factor model fit was improved to be acceptable, $\chi^2(8) = 13.12$, $p = .11$, RMSEA = 0.06, CFI = 0.96, TLI = 0.92.

Second, given the subset of Strange Stories could be classified as cognitive and affective ToM tasks based on the affective component incorporated in each story, a proposed two-factor model was tested. It was hypothesized that Misunderstanding, Double Bluff, and Appearance Reality were indicators that load onto the cognitive ToM factor, while White Lie, Sarcasm, and Persuasion were indicators that load onto affective ToM factor. The proposed model did not fit well , $\chi^2(8) = 16.85$, $p = .03$, RMSEA = 0.07, CFI = 0.93, TLI = 0.87. The fit of the two-factor model was improved by permitting residuals to be correlated between Appearance Reality and White Lie (standardized estimate = .238, $p = .002$). The modified two-factor model provided a good model fit for whole sample, $\chi^2(7) = 7.72$, $p = .36$, RMSEA = 0.02, CFI = 0.99, TLI = 0.99. There was a significant improvement in the modified two-factor model compare with the modified one-factor model, $\Delta\chi^2(1) = 5.40$, $p = .02$. It indicated that the proposed two-factor model was more suitable for the data. Moreover, the modified

two-factor model also had a good fit in both Australian sample, $\chi^2(7) = 7.09$, $p = .42$, RMSEA = 0.01, CFI = 1.00, TLI = 1.00, and mainland Chinese sample, $\chi^2(7) = 4.43$, $p = .73$, RMSEA = 0.00, CFI = 1.00, TLI = 1.07. The standardized loadings of each indicator were all significant in both cultures (ps < .01).

Measurement Invariance of the Two-factor Model of Strange Stories in Mainland Chinese and Australian Sample based on MGCFA

To directly compare children's performance on the cognitive and affective ToM stories between the two samples, the MGCFA was conducted to measure the measurement invariance of the modified two-factor model in both cultures. First, as shown in Table 3, the configural model with both cultures tested simultaneously had a good model fit. It replicated the two latent factors solution holding the same structure in both samples. Second, with constraining factor loadings to be equal across the two cultures, the metric model was found to be well-fitting and had no significant decrease in model fit compared to the configural model, $\Delta\chi^2(4) = 0.63$, $p = .96$. The difference of RMSEA and CFI were both less than .01, and there was even an 0.03 increase of TLI in the metric model.

Table 3. Tests of measurement invariance of the theory of mind latent factor across cultures.

	χ^2	*df*	*p*	$\Delta\chi^2$	Δ *df*	p_{diff}	RMSEA	RMSEA90% CI		CFI	TLI
Model comparison for the whole sample (n =209)											
One-factor	19.16	9	.02				0.07	[0.03,	0.12]	0.92	0.87
Two-factor model	16.85	8	.03	2.31	1	.13	0.07	[0.00,	0.12]	0.93	0.87
Modified one-factor	13.12	8	.11				0.06	[0.00,	0.11]	0.96	0.92
Modified two-factor	7.72	7	.36	5.40	1	.02	0.02	[0.00,	0.09]	0.99	0.99
Modified two-factor model in mainland Chinese (n = 126) and Australian (n = 83) samples											
Mainland China	4.43	7	.73				0.00	[0.00,	0.08]	1.00	1.07
Australia	7.09	7	.42				0.01	[0.00,	0.14]	1.00	1.00
Measurement invariance											
Configural model	11.17	14	.67				0.00	[0.00,	0.08]	1.00	1.05
Metric model	11.80	18	.86	0.63	4	.96	0.00	[0.00,	0.05]	1.00	1.09
Scalar model	22.53	22	.43	10.73	4	.03	0.02	[0.00,	0.08]	1.00	0.99
Partial scalar model	16.86	21	.72	5.06 [a]	3	.17	0.00	[0.00,	0.06]	1.00	1.05
Structural parameter											
Equal latent factor variances model	15.68	23	.87	1.18	2	.55	0.00	[0.00,	0.04]	1.00	1.08
Equal latent factor means model	24.12	25	.51	8.44	2	.01	0.00	[0.00,	0.08]	1.00	1.01
Equal cognitive ToM factor mean model	16.18	24	.88	0.50 [b]	1	.48	0.00	[0.00,	0.04]	1.00	1.08
Equal affective ToM factor mean model	24.02	24	.46	8.34 [c]	1	<.01	0.00	[0.00,	0.08]	1.00	1.00

Note. The modified model means a model with permitting a residual between Appearance Reality and White Lie; a. the partial scalar model was compared to the metric model; b and c. both the equal cognitive ToM factor mean and affective ToM factor mean models were compared to the equal latent factor variances model.

Lastly, the scalar model, in which both factor loadings and intercepts were constrained to be equal across two cultures, showed significant degradation of model fit compared to the metric model $\Delta\chi^2(4) = 10.73$, $p = .03$. There was an 0.02 increase of RMSEA and a 0.10 decrease of TLI. The modification indices suggested that the intercept of Persuasion indicator was noninvariant in the two samples. The gap of the intercepts on Persuasion indicator between the two samples is .31 standardized unit. It suggests that mainland Chinese children found the Persuasion story to be much easier than would be expected on the basis of their performance on the affective ToM latent factor, whereas Australian children found it to be more difficult than would be expected. A possible reason is the utterance "If no one buys the kittens, I'll just have to drown them!" in the Persuasion scenario was a kind of verbal assertion to induce other's guilty feeling. Moreover, guilt induction, which is one common strategy in parental psychological control, was shown to have different influences on children across cultures (Fung & Lau, 2012). In Western culture, parental psychological control such as guilt induction is regarded as negative to children's development because it interferes children's emotional autonomy (Barber & Harmon, 2002). However, East Asian culture emphasizes personal emotion restraint and self-control to accommodate to others in a social group (Matsumoto et al., 2008). Fung and Lau (2012) pointed out that evoking guilt on the parent's perspective empowered 7- to 10-year-old Hong Kong children acquiring empathy to others' feelings. As a result, the current guilt induction Persuasion scenario might be easier for Chinese children because it is similar to their experience of guilt induction from parental psychological control in the cultural context.

As a result, a partial measurement invariance solution was used in the current study to permit unbiased comparison of latent factor means after releasing the equality constrains on the non-invariant parameters. In the partial scalar invariance model, all intercepts except the intercept of Persuasion indicator were set to be equal across the two samples. This model had

a good fit to the data, $\chi^2(21) = 16.86$, $p = .72$, RMSEA = 0.00, CFI = 1.00, TLI = 1.05. There was no significant decrease in model fit compared to the metric model, $\Delta\chi^2(5) = 5.06$, $p = .17$. It indicated that the two-factor model of Strange Stories presented equal form, equal loadings, and equal indicator intercepts except for the intercept of Persuasion indicator across the two cultural groups.

Because the equality of factor loadings and indicator intercepts is the basis of meaningful cross-cultural comparisons of latent factors in the model (Fischer & Karl, 2019), the partial invariance model was used to access the differences in mainland Chinese and Australian children's performance on cognitive and affective ToM stories. First, the variances of the two latent factors were constrained to be equal in the two samples. No significant decrease was found in the model fit indices when comparing with the partial scalar invariance model, $\Delta\chi^2(2) = 1.18$, $p = .55$. It suggests that there was no difference in the within-culture variability of the two factors in Strange Stories between the two samples. Next, the means of the two latent factors were constrained to be equal. There was a significant decrease of model fit, $\Delta\chi^2(2) = 8.44$, $p = .01$, and a 0.07 TLI decrease. In other words, there was a significant cross-cultural difference in the means of the two underlying factors of Strange Stories.

To verify whether the cultural contrast appeared in either one factor or both factors, two more models were administrated. One model set the latent factor mean of cognitive ToM to be equal, and the other model constrained the latent factor mean of affective ToM to be equal across two samples. Then each model was compared to the equal variance model. The result showed that there was no significant change of model fit for the model constraining latent factor mean of cognitive ToM, $\Delta\chi^2(1) = 0.50$, $p = .48$, whereas there was a significant decrease of model fit for the model constraining latent factor mean of affective ToM, $\Delta\chi^2(1) = 8.44$, $p = .004$. It also represented that the latent mean of affective ToM in the Australian sample represented deviations from that of mainland Chinese children when fixing the score

in mainland Chinese sample to be zero. We explored the contrast by post hoc analysis. With fixing the latent mean of affective ToM in mainland Chinese children to be zero, Australian children showed significantly better performance than their mainland Chinese counterparts (Cohen's $d = 0.54$). It indicated a medium to a large difference in performance on affective ToM stories.

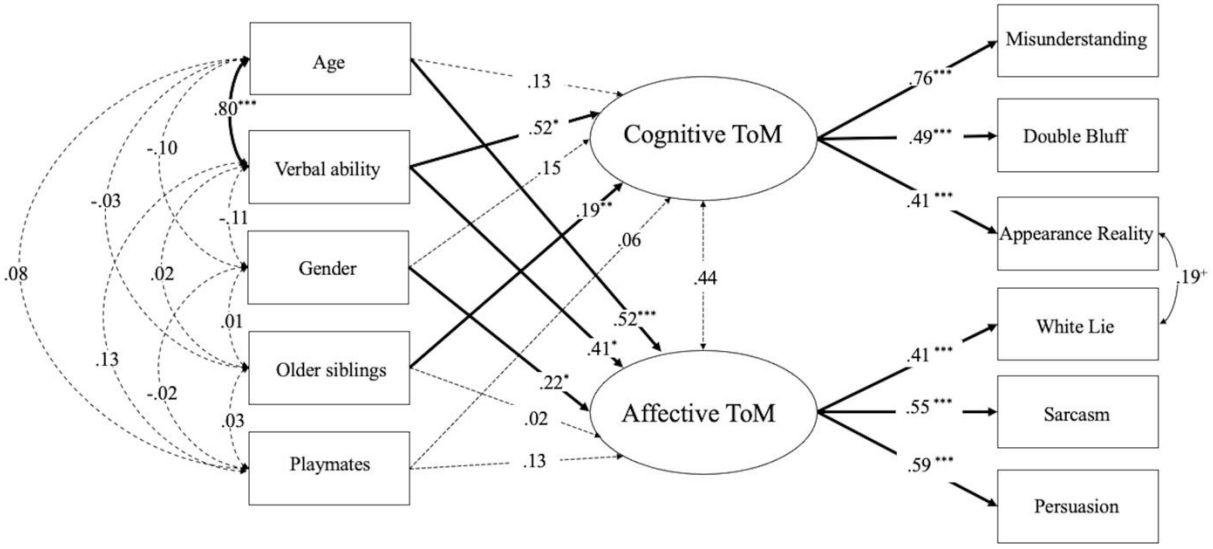
Correlates of Individual Differences in Performance on Cognitive and Affective ToM Stories across Cultures

Another aim of the current study was to investigate the correlates of individual differences in children's performance on cognitive and affective ToM stories in each cultural sample. According to correlation analysis in Table 3, parental education level, the number of younger siblings, and family size did not correlate with the score of any story in both mainland Chinese and Australian samples. These three variables were excluded from the following analysis. The multiple-indicator multiple-causes (MIMIC) model was used for the analysis. The two latent factors of strange stories were regressed by children's age, verbal ability, gender, number of older siblings, and playmates in both cultures with constraining factor loadings, indicator intercepts (except for Persuasion), and variance of latent variables to be equal in the same model. The covariance between each predictor was included in the model to test the unique contribution of each predictor to both cognitive and affective ToM factors. The overall model was acceptable to fit the data, $\chi^2(63) = 66.61$, $p = .35$, RMSEA = 0.02, CFI = 0.99, TLI = 0.98.

For the mainland Chinese sample, the model accounted for significant variances in both children's cognitive ToM ($R^2 = .45$) and affective ToM ($R^2 = .82$). Meanwhile, verbal ability ($B = .27$, $SE = .11$, $z = 2.52$, $p = .01$) and number of older siblings ($B = .22$, $SE = .08$, $z = 2.85$, $p = .004$) were significant predictors of children's scores on cognitive ToM stories, whereas age ($B = .07$, $SE = .02$, $z = 3.55$, $p < .001$), verbal ability ($B = .09$, $SE = .04$, $z = 2.42$, $p = .02$)

and gender (B = .10, SE = .05, z = 2.17, p = .03) were significant predictors of children's scores on affective ToM stories. Moreover, there was a marginal significant residual between Appearance Reality and White Lie (B = .06, SE = .04, z = 1.75, p = .08). Standardized estimates for the model of mainland Chinese sample are presented in Figure 1.

Figure 1. Standardized robust maximum likelihood estimates of correlates of individual differences in cognitive and affective ToM in the mainland Chinese sample.

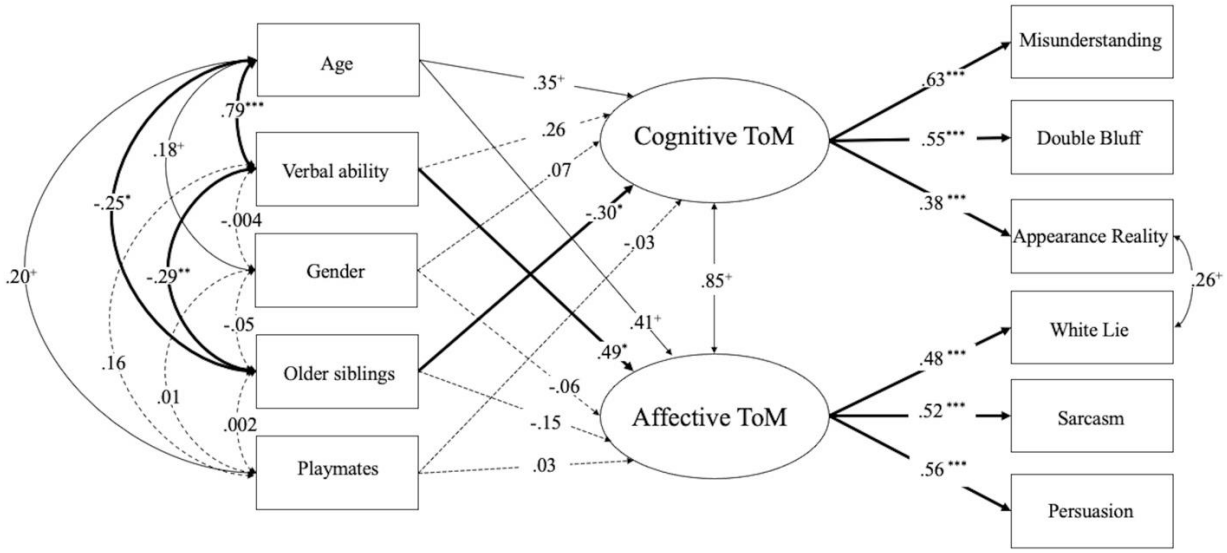


Note. $^{+}$ p <.10, *p < .05; **p < .01; ***p < .001. The standardized PPVT raw scores was used to represent verbal ability in the model.

For the Australian sample, the model also accounted for significant variances in both cognitive ToM (R^2 = .54) and affective ToM (R^2 = .83). Age (B = .11, SE = .07, z = 1.72, p = .09) was marginally significant and number of older siblings (B = -.24, SE = .11, z = -2.16, p = .03) was significant to predict Australian children's performance on cognitive ToM stories, whereas age (B = .05, SE = .03, z = 1.93, p = .05), was marginally significant predictor and verbal ability (B = .11, SE = .05, z = 2.25, p = .02) was significant predictor of Australian children's performance on affective ToM stories. In addition, there was also a marginally

significant residual between Appearance Reality and White Lie (B = .08, SE = .05, z = 1.81, p = .07). Standardized estimates for the model of Australian sample are presented in Figure 2.

Figure 2. Standardized robust maximum likelihood estimates of correlates of individual differences in cognitive and affective ToM in the Australian sample .



Note. $^{+}$ p <.10, *p < .05; **p < .01; ***p < .001. The standardized PPVT raw scores was used to represent verbal ability in the model.

Discussion

Three main findings are apparent in our cross-cultural comparison between children from mainland China and Australia. First, this is the first study which confirmed that Strange Stories fit better and held measurement invariance on a two-factor model rather than a unitary construct across two cultural groups. Second, children in both samples showed similar performance on cognitive ToM stories, whereas Australian children outperformed their mainland Chinese peers on affective ToM stories. Third, cultural differences were found in the factors associated with cognitive and affective ToM.

A Two-factor Model of Strange Stories across Cultures

This is the first study to investigate the measurement invariance of a subset of Strange Stories across cultures. Contrary to the assumption that a unitary construct underpins performance on the Strange Stories task, the current study confirmed a two-latent factor structure of Strange Stories in both mainland Chinese and Australian samples by using MGCFA. Misunderstanding, Double Bluff, and Appearance Reality loaded on the cognitive ToM factor, and these stories required understanding of protagonists' knowledge or cognitive state. In contrast, White Lie, Sarcasm, and Persuasion loaded on the affective ToM factor, and all three scenarios relied on the interpretation of both the characters' cognitive and emotional states. Moreover, the covariance between cognitive and affective ToM factors was not significant in mainland Chinese sample and only marginally significant in Australian sample. This also suggested the cognitive and affective dimensions of ToM are independent and should be treated separately. This is consistent with Wilson et al. (2018)'s findings and the distinction between cognitive and affective processing of ToM (Schurz et al., 2020; Shamay-Tsoory & Aharon-Peretz, 2007).

Cognitive ToM in Mainland Chinese and Australian Children

The findings revealed that Australian and mainland Chinese children presented a similar level of understanding in cognitive ToM stories. The parallel performance was aligned with the identical developmental trajectory of the appreciation of others' false belief understanding in preschoolers from mainland China and Australia (Liu et al., 2008; Shahaeian et al., 2011; Wellman et al., 2001). This is also consistent with Lim et al. (2020)'s finding that 6- to 12-year-old Singapore children did not differ from their Australian peers on second-order false belief task.

We also found the two samples in our study differed in the predictors of their cognitive ToM performance. Age and verbal ability were positively correlated with children's performance on all stories in both samples. However, after including them in the

MIMIC models, the results revealed that while verbal ability was a positive predictor among mainland Chinese sample, age was a marginally significant predictor for Australian children. Mainland Chinese children's age-related differences in cognitive ToM were largely accounted for by their age-related gain in receptive vocabulary as measured by PPVT-R. This may be because these stories are designed based on social situations primarily subsuming interpersonal and verbal interaction (Devine & Hughes, 2016; Ebert, 2020). In contrast, for Australian children, age, not verbal ability, was a positive predictor. These results are not aligned with Wilson et al. (2018)'s finding that verbal ability significantly predicted Australian school-aged children's scores in low affective tone stories when controlling for age and executive functions. A possible explanation is that different verbal ability measures have different predictive utilities in ToM. We only measured Australian children's receptive vocabulary by using PPVT-4 which might not represent all the linguistic requirement involved in Strange Stories. A recent longitudinal study conducted by Ebert (2020) found that receptive vocabulary was less related to German children's advanced ToM measured by Strange stories (two Double Bluff, two Misunderstanding and two White Lie), compared to their receptive grammar/sentence in the longitudinal path models. It indicated that advanced ToM in middle childhood might require more complex language skillset for communication rather than a richer vocabulary.

Furthermore, independent of both age and vernal ability, the number of older siblings positively predicted mainland Chinese children's performance, whereas it was a negative predictor for Australian children. Meanwhile, the number of older siblings was positively correlated with mainland Chinese children's performance on Misunderstanding and Appearance Reality and was negatively related to Australian participants' age, verbal ability and scores on all three cognitive stories. The contrasting role of older siblings might reflect the difference in school-aged children's relationship with older sibling across cultures. Fang

et al. (2003) found that elder siblings were expected to act as role models and provide good examples to younger siblings, and the younger siblings were educated to respect and listen to their older siblings in mainland China. This could be due to Chinese culture emphasizes the respect of authority and in-group harmony (Markus & Kitayama, 1991). As a result, mainland Chinese older siblings take responsibilities to take care of the youngers in the family and they may share their own social experience to educate younger siblings on how to understand others' mental states and maintain the interpersonal relationship. Different from the asymmetric and half-authoritative sibship in mainland China, the sibling relationship between school-aged children in Western countries was featured by equitable negotiation for conflict and symmetry (Fang et al., 2003). Previous studies showed that Western pre-schoolers gain more family-based experiences from older siblings which benefit their ToM (Ruffman et al., 1998). In contrast, during middle childhood, the symmetric and challenging sibship with older siblings might hinder children's social understanding. Additionally, given that the number of older siblings in our Australian sample was negatively related to age and verbal ability, this could increase the possibility for them to become vulnerable and egocentric when interacting with their older siblings.

Affective ToM in Mainland Chinese and Australian Children

In our study, Australian children outperformed their mainland Chinese counterparts on affective ToM stories, and mainland Chinese girls showed superior affective ToM abilities compared to boys. Affective ToM involves children's inferences about others' emotional mental states. The current findings reflect contrasts in the relative salience of emotional mental states across cultures. Chinese preschoolers have been found to exhibit less understanding of emotion knowledge compared to Euro-American children (Wang, 2008). Moreover, mothers' behavioral talk contributes to Chinese preschoolers' ToM (Liu et al., 2016; Lu et al., 2008), whereas mothers' mental state talk is a strong correlate of Western

children's false belief understanding (Adrian et al., 2005; Peterson & Slaughter, 2003). Therefore, the gap in affective ToM might reflect a difference in cultural values whereby in the Chinese culture, individuals' external behaviors and self-control are more highly valued while personal internal mental states are suppressed because they may be socially diverse and do not fit in with the social group (Liu et al., 2016; Markus & Kitayama, 1991).

Furthermore, affective ToM abilities were confirmed to be more advanced in mainland Chinese girls than in boys. It is consistent with previous findings that school-aged girls show an earlier advancement in affective ToM than boys, but no difference in cognitive ToM (Cassetta et al., 2018; Lonigro et al., 2014). The gender discrepancy in mainland Chinese children's affective ToM may also contribute to their poorer performance on affective ToM stories when compared to Australian children. Specifically, there was a positive correlation between gender and scores of White Lie in our Chinese sample, which means girls outscored their male counterparts when interpreting White Lie. Mainland Chinese boys were more likely to provide a rule-based reason, such as "this is polite", without mentioning emotional consequences on the recipients compared to girls. Previous research also revealed that 7- to12-year-old Chinese boys view telling a while lie more positively, and telling a truth more negatively than girls (Ma et al., 2011). This suggests that Chinese boys may simply reason White Lie as a correct/proper expression to show politeness rather than considering its emotion-protection aim.

The cultural differences in affective ToM may also hinge on the use of mental verbs in children's interpretation rather than their comprehension of other's emotional state. Recent research indicated that Chinese preschoolers not only show a similar trend of emotion comprehension as children from Western Europe but also have a better performance on recognizing hidden emotion (Tang et al., 2018). This is consistent with the finding that 7- to 9-year-old children from Iran (characterized by collectivism culture) outperformed their

Australian peers on the Sarcasm scenario (Shahaeian et al., 2014). Although it seems mainland Chinese children should display an equal or even better understanding of others' emotions than their Australian counterparts, our findings were in the opposite direction. This could be due to the requirement of mentioning the exact concept or mental verbs in the coding criteria we adapted from O'Hare et al. (2009). In Shahaeian et al. (2014)'s study, the correct answer of Sarcasm was coded as either mentioning sarcasm explicitly or otherwise presenting a contrast between the literal and hidden meaning in the utterance, such as "she is joking", or "her way of telling him she is upset". However, in our coding guidelines, the former was a full and accurate psychological state response (Score 2) and the latter was treated as a partial psychological state response (Score 1) because it did not contain the exact mental concept "sarcastic/sarcasm". Therefore, we assumed that although our mainland Chinese sample understood the hidden meaning in the affective ToM stories, they may use less mental concept or verbs to explain the characters' responses. Further studies are needed to confirm the origin of the differences in affective stories using culture-appropriate coding guidelines.

We also found that age and verbal ability were both significant predictors of children's affective ToM in the two samples. The significance in verbal ability is in line with previous studies showing an association between verbal abilities and affective ToM in adolescence and young adulthood (Ahmed & Miller, 2011; Vetter et al., 2013). In addition, we found that age uniquely contributed to individual differences in performance on affective ToM stories with verbal ability controlled. This indicates that children's flexibility in simultaneously inferring different mental states and comprehending other's feeling in complex social situations improves with age through middle childhood (Lagattuta et al., 2016; Wilson et al., 2018).

Limitations and Suggestions for Future Research

A limitation of this study is the small sample size of the Australian children. Furthermore, some predictors of children's ToM were only marginally significant, so caution is recommended when drawing conclusions and interpreting the differences. This could be explored in further research by recruiting a large sample size for both cultural groups. The second limitation was that study being conducted at only two sites (mainland China and Australia) for the cultural comparison and found that mainland Chinese children had poorer performance on affective ToM compared to Australian children while there was no difference in cognitive ToM. We interpreted our results in terms of the theoretical framework of collectivism (East) versus individualism (West). However, previous studies demonstrated that the multiplicity of "Eastern" cultures based on delayed false belief understanding of children from Hong Kong and Japan compared to mainland Chinese pre-schoolers (Liu et al., 2008; Naito & Koyama, 2006). Further cross-cultural studies which include three or more samples are needed to form a comprehensive picture of school-aged children's cognitive and affective ToM from different countries. Third, we analysed factors that predicted individual differences of ToM and found some culture-specific results. For example, the number of older siblings were found to positively predict mainland Chinese children's cognitive ToM whereas it was a negative predictor of Australian children's cognitive ToM. It seems that the influence of children's social relationships on their ToM depends on the cultural context. However, other social factors, such as parent-child interaction and peer popularity, were not included in our study. Examination of these variables would be a valuable avenue for future research.

Conclusion

This study confirmed a two-factor model of Strange Stories across cultures and deepened our understanding of cultural differences in school-aged children's cognitive and affective ToM with using Strange Stories. When comparing mainland Chinese children with

Australian children, cultural differences were shown in affective ToM, rather than cognitive ToM. Additionally, factors related to individual differences in ToM across cultures explained the results based on a sociocultural perspective, and provided potential directions for future studies. Lastly, the findings emphasized that both cognitive and affective aspects of ToM should be assessed and considered separately in further research and clinical/educational practices.

Funding

This research was supported by a Griffith University Postgraduate Scholarship.

Conflict of Interest

None.

References

Adrian, J. E., Clemente, R. A., Villanueva, L., & Rieffe, C. (2005). Parent–child picture-book reading, mothers' mental state language and children's theory of mind. *Journal of child language*, *32*(3), 673-686. https://doi.org/10.1017/S0305000905006963

Ahmed, F. S., & Miller, S. L. (2011). Executive function mechanisms of theory of mind. *Journal of autism and developmental disorders*, *41*(5), 667-678. https://doi.org/10.1007/s10803-010-1087-7

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin*, *103*(3), 411-423. https://doi.org/10.1037/0033-2909.103.3.411

Andrews, G., Halford, G. S., Bunch, K. M., Bowden, D., & Jones, T. (2003). Theory of mind and relational complexity. *Child development*, *74*(5), 1476-1499. https://www.jstor.org/stable/3696189

Australian Bureau of Statistics. (2020). Statistics about the population and components of change (births, deaths, migration) for Australia's capital cities and regions. https://www.abs.gov.au/statistics/people/population/regional-population/latest-release#capital-cities

Barber, B. K., & Harmon, E. L. (2002). Violating the self: Parental psychological control of children and adolescents. In B. K. Barber (Ed.), *Intrusive parenting: How psychological control affects children and adolescents* (pp. 15-52). American Psychological Association. https://doi.org/10.1037/10422-002

Berndt, T. J., Cheung, P. C., Lau, S., Hau, K.-T., & Lew, W. J. (1993). Perceptions of parenting in mainland China, Taiwan, and Hong Kong: Sex differences and societal differences. *Developmental Psychology*, *29*(1), 156-164. https://doi.org/10.1037/0012-1649.29.1.156

Bochner, S. (1994). Cross-cultural differences in the self concept: A test of Hofstede's individualism/collectivism distinction. *Journal of cross-cultural psychology*, *25*(2), 273-283. https://doi.org/10.1177/0022022194252007

Bond, M. H., & Cheung, T.-S. (1983). College Students' Spontaneous Self-Concept The Effect of Culture among Respondents in Hong Kong, Japan, and the United States. *Journal of Cross-Cultural Psychology*, *14*(2), 153-171. https://doi.org/10.1177/0022002183014002002

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, *1*(3), 185-216. https://doi.org/10.1177/135910457000100301

Cassetta, B. D., Pexman, P. M., & Goghari, V. M. (2018). Cognitive and affective theory of mind and relations with executive functioning in middle childhood. *Merrill-Palmer Quarterly*, *64*(4), 514-538. https://doi.org/10.13110/merrpalmquar1982.64.4.0514

Cheung, H., Siu, T.-S. C., & Chen, L. (2015). The roles of liar intention, lie content, and theory of mind in children's evaluation of lies. *Journal of Experimental Child Psychology*, *132*, 1-13. https://doi.org/10.1016/j.jecp.2014.12.002

Devine, R. T., & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child development*, *84*(3), 989-1003. https://doi.org/10.1111/cdev.12017

Devine, R. T., & Hughes, C. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the Silent Films and Strange Stories tasks. *Journal of Experimental Child Psychology*, *149*, 23-40. https://doi.org/10.1016/j.jecp.2015.07.011

Dunn, L. M., Dunn, D. M., & Lenhard, A. (2007). *Peabody Picture Vocabulary Test Fourth Edition (PPVT-IV)*. NCS Pearson Minneapolis, MN.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test-revised*. American guidance service, Incorporated.

Ebert, S. (2020). Theory of mind, language, and reading: Developmental relations from early childhood to early adolescence. *Journal of Experimental Child Psychology*, *191*, 104739. https://doi.org/10.1016/j.jecp.2019.104739

Fang, G., Fang, F., Keller, M., Edelstein, W., Kehle, T. J., & Bray, M. A. (2003). Social moral reasoning in Chinese children: A developmental study. *Psychology in the Schools*, *40*(1), 125-138. https://doi.org/10.1002/pits.10074

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in psychology*, *10*, 1507. https://doi.org/10.3389/fpsyg.2019.01507

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*(2), 109-128. https://doi.org/10.1016/0010-0277(95)00692-R

Fung, J., & Lau, A. S. (2012). Tough love or hostile domination? Psychological control and relational induction in cultural context. *Journal of Family Psychology*, *26*(6), 966-975. https://doi.org/10.1037/a0030457

Happe, F. (1994). An adbanced test of theory of mind: understanding of story characoters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129-154. https://doi.org/10.1007/BF02172093

Henan Province Bureau of Statistics. (2019). 2018 Xinyang City National Economic and Social Development Statistical Bulletin (Chinese Website). http://www.xytjj.gov.cn/www/tjzl/tjgb/2020/0429/26107.html

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, *3*(4), 424-453. https://doi.org/10.1037/1082-989X.3.4.424

Hughes, C., Devine, R. T., Ensor, R., Koyasu, M., Mizokawa, A., & Lecce, S. (2014). Lost in translation? Comparing British, Japanese, and Italian children's theory-of-mind performance. *Child Development Research*, *2014*, 893492. https://doi.org/10.1155/2014/893492

Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., Brand, M., Shamay-Tsoory, S., Onur, O. A., & Kessler, J. (2010). Dissociating cognitive from affective theory of mind: a TMS study. *Cortex*, *46*(6), 769-780. https://doi.org/10.1016/j.cortex.2009.07.010

Kennedy, K., Lagattuta, K. H., & Sayfan, L. (2015). Sibling composition, executive function, and children's thinking about mental diversity. *Journal of Experimental Child Psychology*, *132*, 121-139. https://doi.org/10.1016/j.jecp.2014.11.007

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). Guilford Press.

Kobayashi, C., Glover, G. H., & Temple, E. (2007). Cultural and linguistic effects on neural bases of 'Theory of Mind'in American and Japanese children. *Brain Research*, *1164*, 95-107. https://doi.org/10.1016/j.brainres.2007.06.022

Lagattuta, K. H., Elrod, N. M., & Kramer, H. J. (2016). How do thoughts, emotions, and decisions align? A new way to examine theory of mind during middle childhood and beyond. *Journal of Experimental Child Psychology*, *149*, 116-133. https://doi.org/10.1016/j.jecp.2016.01.013

Lecce, S., Bianco, F., Devine, R. T., & Hughes, C. (2017). Relations between theory of mind and executive function in middle childhood: A short-term longitudinal study. *J Exp Child Psychol*, *163*, 69-86. https://doi.org/10.1016/j.jecp.2017.06.011

Lecce, S., Ronchi, L., Del Sette, P., Bischetti, L., & Bambini, V. (2019). Interpreting physical and mental metaphors: Is Theory of Mind associated with pragmatics in middle childhood? *Journal of child language*, *46*(2), 393-407. https://doi.org/10.1017/S030500091800048X

Lewis, C., Huang, Z., & Rooksby, M. (2006). Chinese preschoolers' false belief understanding: Is social knowledge underpinned by parental styles, social interactions or executive function? *Psychologia*, *49*(4), 252-266. https://doi.org/10.2117/psysoc.2006.252

Li, C. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior research methods*, *48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-7

Lim, J., Peterson, C. C., De Rosnay, M., & Slaughter, V. (2020). Children's moral evaluations of prosocial and self-interested lying in relation to age, ToM, cognitive empathy and culture. *European Journal of Developmental Psychology*, *17*(4), 504-526. https://doi.org/10.1080/17405629.2019.1667766

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multiple-group MACS models. In *Modeling contextual effects in longitudinal studies* (pp. 121-147). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of Mind Development in Chinese Children: A Meta-Analysis of False-Belief Understanding Across Cultures and Languages. *Developmental Psychology*, *44*(2), 523-531. https://doi.org/10.1037/0012-1649.44.2.523

Liu, Y., Wang, Y., Luo, R., & Su, Y. (2016). From the external to the internal: Behavior clarifications facilitate theory of mind (ToM) development in Chinese children. *International Journal of Behavioral Development*, *40*(1), 21-30. https://doi.org/10.1177/0165025414562484

Lonigro, A., Laghi, F., Baiocco, R., & Baumgartner, E. (2014). Mind Reading Skills and Empathy: Evidence for Nice and Nasty ToM Behaviours in School-Aged Children [Article]. *Journal of Child and Family Studies*, *23*(3), 581-590. https://doi.org/10.1007/s10826-013-9722-5

Lu, H., Su, Y., & Wang, Q. (2008). Talking about others facilitates theory of mind in Chinese preschoolers. *Developmental Psychology*, *44*(6), 1726-1736. https://doi.org/10.1037/a0013074

Ma, F., Xu, F., Heyman, G. D., & Lee, K. (2011). Chinese children's evaluations of white lies: Weighing the consequences for recipients. *Journal of Experimental Child Psychology*, *108*(2), 308-321. https://doi.org/10.1016/j.jecp.2010.08.015

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological review*, *98*(2), 224-253. https://doi.org/10.1037/0033-295X.98.2.224

Matsumoto, D., Yoo, S. H., & Nakagawa, S. (2008). Culture, emotion regulation, and adjustment. *Journal of Personality and Social Psychology*, *94*(6), 925-937. https://doi.org/10.1037/0022-3514.94.6.925

McAlister, A., & Peterson, C. (2007). A longitudinal study of child siblings and theory of mind development. *Cognitive Development*, *22*(2), 258-270. https://doi.org/10.1016/j.cogdev.2006.10.009

Miller, S. A. (2013). Children's understanding of second‐order false belief: Comparisons of content and method of assessment. *Infant and Child Development*, *22*(6), 649-658. https://doi.org/10.1002/icd.1810

Naito, M., & Koyama, K. (2006). The development of false-belief understanding in Japanese children: Delay and difference? *International Journal of Behavioral Development*, *30*(4), 290-304. https://doi.org/10.1177/0165025406063622

O'Hare, A. E., Bremner, L., Nash, M., Happé, F., & Pettigrew, L. M. (2009). A clinical assessment tool for advanced theory of mind performance in 5 to 12 year olds. *Journal of autism and developmental disorders*, *39*(6), 916-928. https://doi.org/10.1007/s10803-009-0699-2

Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: You catch it from your sibs. *Child development*, *65*(4), 1228-1238. https://doi.org/10.2307/1131316

Peterson, C. C. (2000). Kindred spirits: Influences of siblings' perspectives on theory of mind. *Cognitive Development*, *15*(4), 435-455. https://doi.org/10.1016/S0885-2014(01)00040-5

Peterson, C. C., & Slaughter, V. (2003). Opening windows into the mind: Mothers' preferences for mental state explanations and children's theory of mind. *Cognitive Development*, *18*(3), 399-429. https://doi.org/10.1016/S0885-2014(03)00041-8

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(04), 515-526. https://doi.org/10.1017/S0140525X00076512

Prime, H., Plamondon, A., Pauker, S., Perlman, M., & Jenkins, J. M. (2016). Sibling cognitive sensitivity as a moderator of the relationship between sibship size and children's theory of mind: A longitudinal analysis. *Cognitive Development*, *39*, 93-102. https://doi.org/10.1016/j.cogdev.2016.03.005

Ruffman, T., Perner, J., Naito, M., Parkin, L., & Clements, W. A. (1998). Older (but not younger) siblings facilitate false belief understanding. *Developmental psychology*, *34*(1), 161-174. https://doi.org/10.1037/0012-1649.34.1.161

Sang, B., & Miao, X. (1990). The revision of trail norm of Peabody picture vocabulary test revised (PPVT-R) In Shanghai proper. *Psychological Science (Chinese Journal)*, *5*, 20-25.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2020). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*. https://doi.org/10.1037/bul0000303

Shahaeian, A., Nielsen, M., Peterson, C. C., & Slaughter, V. (2014). Cultural and family influences on children's theory of mind development: A comparison of Australian and Iranian school-age children. *Journal of Cross-Cultural Psychology*, *45*(4), 555. https://doi.org/10.1177/0022022113513921

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. https://doi.org/10.1037/a0023899

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. *Neuropsychologia*, *45*(13), 3054-3067. https://doi.org/10.1016/j.neuropsychologia.2007.05.021

Shamay-Tsoory, S. G., Harari, H., Aharon-Peretz, J., & Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex*, *46*(5), 668-677. https://doi.org/10.1016/j.cortex.2009.04.008

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social neuroscience*, *1*(3-4), 149-166. https://doi.org/10.1080/17470910600985589

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., Goldsher, D., & Aharon-Peretz, J. (2005). Impaired "affective theory of mind" is associated with right ventromedial prefrontal damage. *Cognitive and Behavioral Neurology*, *18*(1), 55-67. https://doi.org/10.1097/01.wnn.0000152228.90129.99

Slaughter, V., & Perez-Zapata, D. (2014). Cultural variations in the development of mind reading. *Child Development Perspectives*, *8*(4), 237-241. https://doi.org/10.1111/cdep.12091

Tang, Y., Harris, P. L., Pons, F., Zou, H., Zhang, W., & Xu, Q. (2018). The understanding of emotion among young Chinese children. *International Journal of Behavioral Development*, *42*(5), 512-517. https://doi.org/10.1177/0165025417741366

Tardif, T., & Chan, C. (2005). Comprehension and production of nouns and verbs: Data from CDI norming studies in English, Mandarin, and Cantonese. 10th International Congress for the Study of Child Language, Berlin, Germany,

Triandis, H. C. (2001). Individualism-collectivism and personality. *Journal of Personality*, *69*(6), 907-924. https://doi.org/10.1111/1467-6494.696169

Vetter, N. C., Altgassen, M., Phillips, L., Mahy, C. E., & Kliegel, M. (2013). Development of affective theory of mind across adolescence: Disentangling the role of executive functions. *Developmental Neuropsychology*, *38*(2), 114-125. https://doi.org/10.1080/87565641.2012.733786

Wang, Q. (2008). Emotion knowledge and autobiographical memory across the preschool years: A cross-cultural longitudinal investigation. *Cognition*, *108*(1), 117-135. https://doi.org/10.1016/j.cognition.2008.02.002

Wang, Z., Devine, R. T., Wong, K. K., & Hughes, C. (2016). Theory of mind and executive function during middle childhood across cultures. *Journal of Experimental Child Psychology*, *149*, 6-22. https://doi.org/10.1016/j.jecp.2015.09.028

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

Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theroy-of-mind understanding in Chinese children. *Psychological Science*, *17*(12), 1075-1081. https://doi.org/10.1111/j.1467-9280.2006.01830.x

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. https://doi.org/10.1111/j.1467-8624.2011.01583.x

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child development*, *80*(4), 1097-1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x.

Wilson, J., Andrews, G., Hogan, C., Wang, S., & Shum, D. H. (2018). Executive function in middle childhood and the relationship with theory of mind. *Developmental neuropsychology*, *43*(3), 163-182. https://doi.org/10.1080/87565641.2018.1440296

Supplementary materials.

Details of Strange Stories.

White Lie. White Lie is the scenario that people tell lies to protect other's feelings because the blunt truth is hurtful or impolite (Cheung et al., 2015). In the story, a boy named Peter loves his aunt very much, but does not like his aunt's new hat. When his aunt asks him, "How do you like my new hat?", he answers, "Oh, it's very nice". Then two questions were asked, 1. "Was it true what Peter said?" and 2. "Why did he say it?" The while lie scenario refers to the understanding of other's emotional state, which is "A **knows** B will get **sad** if A does something".

Sarcasm. Sarcasm is one kind of ironic speech used to express implicit criticism and negative feelings (Shamay-Tsoory et al., 2005). In the story, Sarah and Tom are going on a picnic. It is Tom's idea. He says it will be a lovely sunny day. But when they get the food out, it rains, and the food gets all wet. Sarah says: "Oh yes, a lovely day for a picnic alright!" Then children were asked two questions, 1. "Is it true what Sarah says?' and 2. "Why does she say this?". To understand sarcasm, children first need to infer that Sarah knows it is not a lovely day and then to infer that she is unhappy now.

Persuasion. Persuasion is the process of directing a person to the adoption of a belief, an attitude, or an idea by a communicative mean. In the story, Jill planned to buy a kitten from Mrs Smith. Mrs Smith loved kittens and wouldn't harm them although she couldn't not keep so many kittens by herself. However, Jill wasn't sure she wanted to buy one kitten, because they were all males but she wanted a female. Then Mrs Smith said, ''If no one buys the kittens, I'll just have to drown them!'' Then children were asked two questions, 1."Was it true what Mrs Smith said? " and 2. "Why did Mrs Smith say this to Jill?" To interpret Mrs Smith's intention, participants were required to know that her aim was not to drown the kittens, but to induce Jill's guilty feeling for the kittens.

Misunderstanding. In the story, a burglar who has just robbed a shop dropped his glove. A policeman on his beat sees him drop his glove, but doesn't know the man is a burglar. He just wants to tell him he dropped his glove, then he shouts out to the burglar, ''Hey you, Stop!''. The burglar sees the policeman, puts his hands up and admits that he did the break-in at the local shop. Then participants were asked two questions, 1. "Was the policeman surprised by what the burglar did? " and 2. "Why did the burglar do this, when the policeman just wanted to give him back his glove?" To explain the burglar's reaction, participants needed to make recursive reasoning which is "the burglar thinks the police knows he has just robbed a shop".

Double Bluff. In the story, Jim has a brother named Simon, who always lies. And Jim knows that Simon never tells the truth. One day Simon stole Jim's ping-pong bat. Jim can't find it and asked Simon, ''Where is my ping- pong bat? You must have hidden it either in the cupboard or under your bed because I've looked everywhere else. Where is it, in the cupboard or under your bed?'' Simon tells him the bat is under his bed. Then participants were asked three questions, 1. "Was it true what Simon told Jim?", 2. "Where will Jim look for his ping-pong bat?" and 3." Why will Jim look there for his bat? " To provide the correct answer (the cupboard), participants needed to know that, "Jim thinks Simon will lie to him."

Appearance Reality. The Appearance Reality task requires the children to know that the appearance of an object does not necessarily correspond to its reality. In the story, Alice sees his neighbour Mr. Brown is dressed up as Santa Claus, giving out sweets to all the children in the store. Alice thinks she recognises Mr Brown, so asks him ''Who are you?'' Mr Brown answers ''I'm Santa Claus!'' Then children were asked two questions, 1. "Is it true what Mr Brown says? " and 2. "Why does he say this?" To correctly answer the question, children needed first to know Mr Brown just pretends to be Santa, and then to infer that he says he is Santa because he wants Alice to think he is Santa.

Table S1. Correlations among key study variables in the whole sample (n = 209)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Age	-													
2. Verbal ability	.72***	-												
3. Gender	.02	-.03	-											
4. Parental EDU	-.03	.20**	-.04	-										
5. Older sibling	-.15*	.01	.01	-.02	-									
6. Younger sibling	.15*	.20**	.13+	.13+	-.26***	-								
7. Playmates	.15*	.13+	.00	.00	.00	-.08	-							
8. Family size	-.06	-.06	.09	.01	.14*	.45***	.03	-						
9. MU	.43***	.35***	.09	-.02	-.05	.02	.07	.01	-					
10. Double Bluff	.27***	.35***	.05	.01	-.21**	.03	.04	-.02	.32***	-				
11. AR	.20**	.29***	.07	.10	.04	.03	.06	-.03	.31***	.20**	-			
12. White Lie	.33***	.41***	.18*	.05	-.09	.06	.05	-.07	.26***	.19**	.32***	-		
13. Sarcasm	.46***	.44***	-.05	.10	-.03	.04	.23***	-.01	.32***	.25***	.09	.19**	-	
14. Persuasion	.49***	.43***	.05	-.04	-.11	.00	.07	-.12+	.27***	.21**	.17*	.30***	.33***	-

Note. $^{+}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. Gender was coded as: 0 = male, 1= female. Parental EDU = Parental education level, MU = Misunderstanding, AR = Appearance Reality.