Title	Development and psychometric properties of a culturally adapted video version of strange stories as a measure of advanced theory of mind in youths
Author(s)	Yong-Hwee Nah and Mo Chen

This is the accepted author's manuscript of the following article:

Nah, Y.-H., & Chen, M. (2022). Development and psychometric properties of a culturally adapted video version of strange stories as a measure of advanced theory of mind in youths. *Journal of Psychoeducational Assessment*. Advance online publication. https://doi.org/10.1177/07342829221075981

Running Head: Y-ToM

Development and Psychometric Properties of a Culturally Adapted Video Version of Strange Stories as a Measure of Advanced Theory of Mind in Youths

Yong-Hwee Nah and Mo Chen

National Institute of Education, Nanyang Technological University, Singapore.

Correspondence concerning this article should be addressed to Yong-Hwee Nah, National Institute of Education NIE2-03-106, 1 Nanyang Walk, Singapore 637616, Nanyang Technological University, Singapore 637616. Email: yonghwee.nah@nie.edu.sg

ORCID: 0000-0003-3748-5710

Mo Chen, National Institute of Education NIE5-B3-32, 1 Nanyang Walk, Singapore 637616, Nanyang Technological University, Singapore 637616. Email: mo.chen@nie.edu.sg

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements

This study was funded by the Education Research Funding Programme, National Institute of Education (NIE), Nanyang Technological University, Singapore, project no. OER 31/17 NYH. The views expressed in this paper are the author's and do not necessarily represent the views of NIE.

The authors would like to thank all the participants in this study. The author acknowledged the invaluable contribution of Hillary Lim, research assistant in this study.

Abstract

This study described the development of a culturally adapted video version of Strange Stories test as a measure of advanced theory of mind for youths in an Asian country (i.e., Singapore), the Y-ToM, and to provide preliminary psychometric properties. Participants were 170 youths (82 male, 88 female) aged from 13 – 16 years old (M = 14.77, *SD* = 1.16) in Singapore. The youths completed the Y-ToM, an abbreviated IQ test and the Happé's Strange Stories in a counterbalanced order while their parents completed the Child Behavior Checklist (CBCL). A two-factor structure consisting of social and physical subscales was suggested. Concurrent, convergent, divergent, and diagnostic validity of the Y-ToM was examined. Internal consistency of the Y-ToM social subscale was acceptable though it was not satisfactory for the Y-ToM physical subscale. Inter-rater reliability was good while test-retest reliability was lower.

Key words: Advanced Theory of Mind; Youths; Asian; Psychometric properties; Strange Stories

Development and Psychometric Properties of a Culturally Adapted Video Version of Strange Stories as a Measure of Advanced Theory of Mind in Youths

Having a Theory of Mind (ToM) allows us to recognise other people's mental states, predict and explain people's behaviour (Premack & Woodruff, 1978). Deficits in ToM affect individuals' ability to interact appropriately in a particular social situation and have been suggested to account for the social cognitive deficits (such as the difficulty to take the perspective of others) observed in some individuals with social difficulties such as Autism Spectrum Disorder (ASD) (White et al., 2009).

To date, much research has been done to examine ToM development in early and middle childhood, and only a few studies have examined development of ToM in youths (Meinhardt-Injac et al., 2020). The ability to make sense of social interactions/situations is especially vital for youths. Youths with issue or difficulty in social interactions tend to face negative consequences such as rejection by peers (Wolters et al., 2014) and difficulty in maintaining reciprocal friendship (Salisch et al., 2014). Deficits in ToM were also suggested to be a risk factor for naïve involvement in criminal activities (cf. Brewer & Young, 2015). Identifying those potential challenges and fostering effective socio-emotional learning is therefore pertinent to improving youths' social and behavioural competence, especially those at risk of developing socio-emotional difficulties.

Due to the ceiling effect of typical (e.g., false belief) ToM tests, over the years, there has been an increased focus to measure advanced ToM beyond childhood years. Advanced ToM refers to one's ability to recognise others' intentions, attitudes, desires and sometimes emotions and not just false belief (Happé, 1994). Currently, measures available to examine advanced ToM have been developed (e.g., the Reading the Mind in the Eyes Test; Baron-Cohen et al., 2001, Adult-Theory of Mind; Brewer et al., 2017, the Movie for the Assessment of Social Cognition; Dziobek et al., 2006, and the Strange Stories test; Happé, 1994).

For instance, the Strange Stories test requires examinees to explain the meaning of the behaviour of the key characters within the scenarios when they use expressions that mean something other than what a literal interpretation of the expression might suggest (e.g., figure of speech, sarcasm, white lies, etc.). However, the Strange Stories task is typically presented in a written format that does not examine one's ability to process naturalistic social cues such as facial expression and vocal intonation in real time (Murray et al., 2017). There is also the need to assess subtle ToM difficulties in an ecologically valid manner where individuals are exposed to the fast-paced nature of real-life social interaction.

To address the above concerns, the current study sought to develop a culturally adapted video version of Strange Stories test as a measure of advanced ToM targeted at youths in an Asian country (i.e., Singapore), the Y-ToM. The Y-ToM is named to refer to a youth version of ToM test adapted from the mental state constructs and scenarios based on Happé's Strange Stories test (Happé, 1994). The adoption of the video-based format echoes the argument that video-based scenarios are more ecologically valid, because they require individuals to make inferences based on more authentic, dynamic, and various socially relevant cues (Livingston et al., 2019). Moreover, using video-based presentations of ToM items has been reported to be more sensitive in detecting the nuanced impairments in ToM in adults with autism compared typically developing controls (e.g., Murray et al., 2017).

The Strange Stories test has been adapted as a Farsi (paper-and-pen format) translation for an Iranian community sample of school-aged children aged 9-11 (Shahrivar et al., 2017) and also been adapted as a video version to assess ToM in Australian sample of adults (mean age of 27 years old) with ASD (Brewer et al., 2017). Expanding available advanced ToM tools in an Asian context could be a good reason to adapt the Strange Stories to understand the development of advanced ToM in the light of cultural differences (Perez-Zapata et al., 2016; Slaughter & Perez-Zapata, 2014). For instance, previous research

suggests that there may be differences in cultural aspects of understanding mental states such as liars' intentions (Cameron et al., 2012; Cheung et al., 2015). A study found that their Iranian school-aged sample understood the concept of white lie better than English-speaking children (Shahrivar et al., 2017). Other research studies also found that Chinese children perceived white lie as a pro-social behaviour and they understood that white lies might not be negative as compared to their European counterparts (Cameron et al., 2012; Ma et al., 2011). Another study found that people obtained higher scores on ToM tasks when considering other people from the same cultural background as compared to when considering people from another culture (Perez-Zapata et al., 2016).

In summary, this study aimed to describe the development of the Y-ToM and to examine its psychometric properties in a community sample of Singaporean youths with ages from 13 to 16 years old. This age group was chosen because they would typically be youths in secondary schools based on the Singapore's formal education system and so they would be exposed to similar social, cultural and educational context. Specifically the Y-ToM used videos as the presentation mode to examine how youths recognise others' intentions and desires (mental state attribution) in a more naturalistic way in a Singaporean sample. It is also important to create a localised measurement tool given that people from the same cultural background often share a common understanding of what is relevant in a given situation, which in turn helps to promote accurate understanding of social situations or interactions (Apperly, 2010).

In this study, we also examined the youths' emotional and behavioural problems as rated by their parents on the Child Behaviour Checklist (CBCL; Rescorla et al., 2007) and their possible associations with the Y-ToM score. The CBCL parent-rated form is a commonly used screening tool and has shown criterion validity in a large Singapore sample (Ang et al., 2012) and the CBCL has also shown to be effective in identifying children with

social difficulties such as ASD in Singapore (Ooi et al., 2011). Based on the findings from Ooi et al. (2011)'s study on the use of CBCL for screening children aged between 4 and 18 years (mean age = 9.06) suspected of having ASD in Singapore, the Withdrawn/Depressed, Social problems, Thought problems, Aggressive behavior and Attention problems subscales significantly discriminated the ASD group from the non-ASD groups while other CBCL subscales such as Somatic problems was less discriminative. Given that ToM has been implicated in the social cognitive deficits typically observed in some individuals with social difficulties such as ASD (White et al., 2009), therefore, to establish convergent and divergent validity in this study, we expected that the Y-ToM social score would be correlated to a certain degree with the Withdrawn/Depressed, Social problems, Thought problems, Aggressive behavior and Attention problems subscales raw scores and to a lesser degree with Somatic problems subscale raw score.

Using the COnsensus-based Standards for the selection of health status Measurement Instruments (COSMIN) checklist of measurement properties (Mokkink et al., 2010) as a framework, this study aims to establish the validity evidence (i.e., content validity, construct validity, and criterion validity) for the Y-ToM, in addition to examining its reliability. Four specific research questions are as follows:

1. Do the items of the Y-ToM have content validity?
2. Does the Y-ToM have sufficient evidence for construct validity?
    a. Does the Y-ToM demonstrate a two-factor (i.e., one social factor and one physical factor) structure? (structural validity)?
    b. After accounting for age, sex, verbal IQ, and nonverbal IQ scores, does the social latent factor of the Y-ToM correlate with the social latent factor of the Happe's original Strange Stories? (convergent validity)?
    c. After accounting for age, sex, verbal IQ, and nonverbal IQ scores, does the

social latent factor of the Y-ToM correlate more with some subscales of CBCL, but less with some other subscales of CBCL? (convergent and divergent validity)?

   d. Do participants who score higher on CBCL tend to have higher Y-ToM social scores? (diagnostic validity)?
   e. Do females score higher on the Y-ToM social score as compared to males but there are no significant difference on the Y-ToM physical score (hypothesis testing as part of construct validity)?

3. Does the Y-ToM have sufficient evidence for criterion validity? That is, to what degree does the performance on the Y-ToM correlate with that on the Happe's original Strange Stories? (concurrent validity as a type of criterion-related validity)?
4. Does the Y-ToM have sufficient reliability (i.e., internal consistency, test-retest reliability, and inter-rater reliability)?

Method

Participants

This study was approved by the Institutional Review Board of XXX. All participants were recruited via general advertising in mass and social media. A total number of 170 youths (82 male, 88 female) with no known/identified physical or/and developmental disabilities studying in regular secondary schools in Singapore participated in this study. The numbers and percentages of the participants at Secondary One (aged 13), Secondary Two (aged 14), Secondary Three (aged 15) and Secondary Four (aged 16) was 49 (28.8%), 49 (28.8%), 39 (22.9%), and 33 (19.4%) respectively. During the testing, there were six youths whose parents were concerned of their children having some form of learning difficulties. However, we decided to include these six youths in our evaluation sample (as part of

collecting validity evidence) as they were reported to have no significant issues in schools and they did not differ from the other participants in terms of IQ scores, internalising and externalising problems as reported by parents.

The mean age of all participants was 14.77 years (*SD* = 1.16, range: 13.01-16.98 years), and did not differ significantly by sex. In terms of ethnicity, the group comprised 81.2% Chinese, 4.1% Malay, 8.8% Indian, and 5.9% Others. Participants were excluded if their overall IQ score fell below 85 [based on the Wechsler Abbreviated Scale of Intelligence – Second Edition (WASI-II; Wechsler, 2011)] and all participants in this study had an overall IQ score above 85. The mean verbal, non-verbal, and overall IQ of all participants were 110.24 (*SD* = 9.82, range = 82-132), 118.74 (*SD* = 12.44, range = 83-152), and 116.15 (*SD* = 10.14, range = 85-139) respectively.

The parents of the youth participants (*N* = 170, mean age = 45.09 years, *SD* = 4.09, range = 32-57) were also recruited to complete a basic demographic form and a parent-rated form on their child's behaviours using the Child Behaviour Checklist (CBCL; Rescorla et al., 2007). Majority of the parents had a degree (45.3%) followed by a postgraduate degree (27.1%).

Measures

Development of the Y-ToM: A Singapore Adapted Strange Stories

Scripts for the Y-ToM were developed from the first author's clinical experience, research literature and inputs from the research team. The Y-ToM consisted of two separate set of items: one comprising of 14 social items and the other including 8 physical items. The social items tested the examinee's ability to attribute a speaker's intention from seven categories including: (1) lie; (2) white lie; (3) misunderstanding; (4) double bluff; (5) figure of speech; (6) sarcasm; and (7) persuasion. There were two items for each category. Four secondary school teachers were asked to examine whether the scenarios in social items were

typical of how secondary school students behave (face validity) and amendments to the scripts were made accordingly. The Y-ToM physical items tried to follow as closely as possible to the content of Happé's original stories but were modified to fit the local context and age group. For instance, in Happé's original physical story of 'X-ray' where an old lady slipped on her icy door step and had to go take an X-ray, we changed it to a youth playing basketball and falling down while playing. The physical items required logical reasoning to decipher the characters' utterances or behaviours.

The language used in the scripts followed as closely to everyday spoken language as possible. Some of the modifications made in the Y-ToM items reflected Singapore's cultural phenomena such as the common usage of instant messaging in our youths (in some of our videos) and how concepts (e.g., the concept of installment and interest rate) were taught in the Singapore education system.

Six actors (undergraduate students) were recruited via personal contacts and general advertising on campus. Preference was given to those who had some acting background and no formal training was provided to these actors. A third-year undergraduate majoring in Communication Studies was hired to produce the videos. English captions were provided for each video.

Pilot Testing. The 22 videos (with the corresponding questions) were pilot tested and edits to the videos were made based on the participants' feedback. More information on the pilot testing phase is provided under the Content Validity of the Results section and in Online Supporting Material 1. Based on the timing, we set the time limit to 60 seconds (using mean + 1.5 *SD*, i.e., 45.98 + 13.2) for the actual evaluation sample. When the question appeared on the screen and once 60 seconds had elapsed, the question page would automatically advance to the next video and participants would not be able to continue typing the answers. The final Y-ToM videos ranged in duration from 15 to 52 seconds. Participants' answers were scored

on a 0-2 scale: 0 (incorrect), 1 (partially correct) or 2 (correct). The score range for the social subscale is from 0 to 28 while the physical subscale is from 0 to 16. An example of both the scripts of Y-ToM social and physical items was provided in Appendix A. The complete set of videos can be accessed via the link below:

Social playlist link:
https://www.youtube.com/watch?v=re6i8u1R4_Y&list=PL4NPNFHEd_BEQXvsLWNPo6pl2xP_yukrK

Physical playlist link:
https://www.youtube.com/watch?v=abqW8gHipPE&list=PL4NPNFHEd_BEcovCqtmUMY-ftvGTcdZri

Happé's Strange Stories

The Strange Stories test (written version) was used as a concurrent validity measure of the Y-ToM. In this study, our version consisted of 10 social and 8 physical stories that were used in O'Hare et al. (2009) and White et al. (2009). In this study, the Strange Stories were presented as an online survey version in Qualtrics via a Samsung Galaxy Tab S3 tablet. Similar to the Y-ToM, we created four different versions with social and physical scenarios distributed randomly throughout each version for counterbalancing purposes. The questions relating to each scenario were displayed on the tablet screen following the story, and participants were asked to type out the answer using the keyboard provided. Participants were instructed that there was no time limit within which they had to respond. Participants' answers were scored on a 0-2 scale: 0 (incorrect), 1 (partially correct) or 2 (correct) using the scoring criteria reported in O'Hare et al. (2009) and White et al. (2009).

IQ Measure

Participants were administered four subtests of the WASI-II (Wechsler, 2011): Block Design, Vocabulary, Matrix Reasoning, and Similarities. The WASI-II is an abbreviated

measure of cognitive intelligence designed for individuals aged 6 to 90 years old and can be administered in no more than 30 minutes. The Perceptual Reasoning (nonverbal) component comprised of the Block Design and Matrix Reasoning subtests, while the Verbal Comprehension component comprised of the Vocabulary and Similarities subtests. Composite scores were calculated from these subscales to create a Perceptual Reasoning Index (PRI), a Verbal Comprehension Index (VCI), and a Full Scale Intelligence Quotient (FSIQ). Reliability and validity data of the WASI-II were reported in McCrimmon and Smith (2013).

Child Behavior Checklist

The Child Behaviour Checklist (CBCL) is a commonly used dimensional parent-rated tool to screen for emotional and behavioural problems in children and adolescents (Rescorla et al., 2007). The CBCL produces scores on eight narrow subscales (e.g., Withdrawn, Somatic complaints, etc.) and three broadband scales (i.e., Internalising Problems, Externalising Problems, and Total Problems). The CBCL parent-rated form (6-18 years old) were used in this study. The form includes 118 items that describe the child's behavioural, emotional, and social problems over the past 6 months. These items are rated on a 3-point scale (0 = Not True [as far as you know), 1 = Sometimes or Somewhat True, or 2 = Very True or Often True). Reliability and validity data of the CBCL were reported in Achenbach and Rescorla (2001).

In this study, the CBCL was completed by parents to evaluate their possible associations with ToM ability, specifically on the Y-ToM social score. It should be noted that raw scores were used in the analyses as recommended by Achenbach and Rescorla (2001) as this current study involved typically developing individuals as compared to using clinical samples.

Procedure

Parents' consent and child's assent were obtained prior to data collection. The parents completed the demographic form as well as the CBCL. To reduce order effect, the Y-ToM and Strange Stories administration were counterbalanced for half of the sample where one half did the Y-ToM first, followed by the WASI-II, and then the Strange Stories while the other half did the Strange Stories first, followed by the WASI-II, and then the Y-ToM. There was always a time gap (of at least 30 minutes) between the administration of the Y-ToM and the Strange Stories (and vice versa) to minimise practice or familiarity effect. The whole testing session took about two hours to complete, including the break time.

During the testing session, the research assistant administered the Y-ToM/Strange Stories online (using Qualtrics) to the participants in their homes. After the participants had completed the Y-ToM/Strange Stories, the research assistant then administered the WASI-II. After the IQ administration, the participants took a short break (10 minutes) before the Y-ToM/Strange Stories was administered to them. To evaluate test-retest reliability, the Y-ToM was administered again for 26.5% (N = 45) of the sample, following at least a 28-day gap (the mean interval duration was 42.11, SD = 17.72, range = 28-139) from their initial assessment. Inter-rater reliability was also evaluated with two raters scoring the Y-ToM items using 91.2% (N = 155) of the sample. Each dyad of participant (both adolescent and parent) was given monetary vouchers equivalent to a total of SGD $25 as a token of participation upon their completion. For those youths who were randomly selected for the test-retest administration of the Y-ToM, they received an additional monetary voucher equivalent to SGD $5 as a token of participation.

Data Analysis

Descriptive statistical analysis of participants' performance on the Y-ToM was performed using the SPSS (release 26). Structural validity (as part of construct validity) was established using the confirmatory factor analysis (CFA) in Mplus 8 (Muthén & Muthén,

2017). Concurrent, convergent, and divergent validity were assessed using the Y-ToM scores with Happé's Strange Stories, age, sex, IQ scores, and CBCL scores using Structural Equation Modeling (SEM) in Mplus 8. Based on prior published criteria by Hu and Bentler (1999), Kline (2015), and McDonald (2013), the following standards for 'good fit' were adopted: Comparative Fit Index (CFI) and Tucker Lewis Index (TFI) > .95, Root Mean Square Error of Approximation (RMSEA) < .06, and the WRMR/SRMR value <1.0. Diagnostic validity was examined by comparing the Y-ToM social score with the CBCL Social Problem's T-score cutoff of 60 (1 standard deviation above the mean) using the ANOVA. ANOVA was also used to compare the performance of the Y-ToM scores based on sex. Lastly, Cronbach's alpha coefficient via SPSS was used to find the internal consistency and split-half reliability of the Y-ToM social and physical scales. Test-retest and inter-rater reliability were calculated using intra-class correlation in SPSS as well.

Results

Descriptive Statistics

There were no missing data reported. Table 1 (Online Supporting Material 2) summarised the descriptive data of the youths' performance on each of the story type of the Y-ToM. The youth participants struggled the most (score of 0) for the Social sub-category of Persuasion (Pocket money) (27%) followed by Deception (Bake sale) (22%) while they found White lie (83-86%) and Figure of speech (Melting; 86%) the easiest (score of 2). In terms of floor and ceiling on the Y-ToM score, 0.6% (N = 1) and 2.9% (N = 5) obtained the maximum score on the social (maximum score of 28) and physical (maximum score of 16) total score respectively while 2.4% (N = 4) and 0.6% (N = 1) obtained the lowest score on the social (score of 12) and physical (score of 5) total score respectively.

Evidence for Content Validity and Construct Validity

Content Validity

Detailed information about the pilot testing phase is available as Online Supporting Material 1. To summarise, the 22 videos (with the corresponding questions) were presented as an online survey version via Qualtrics, and pilot tested with 20 youths (9 male, 11 female; mean age = 14.95, range: 13-16 years old) studying in local secondary schools. The main aim of the pilot test was to ensure that the content and language used in the videos were suitable for the target sample (youths between 13 to 16 years old), and also to determine what time limit to respond to the question was reasonable and appropriate for most of the pilot participants. Descriptive data of the pilot Y-ToM are available in the Online Supporting Material 1. Edits to the videos were made based on the participants' feedback.

Structural Validity

Using the CFA, the 22 items of the Y-ToM were tested against the hypothesised two-factor (social and physical based on Happé's Strange Stories) structure with estimations based on the weighted least square mean and variance adjusted (WLSMV) function. This model provided an acceptable to good fit to the data, $\chi^2(208) = 223.04$, $p = .23$, CFI = .94, TLI = .93, RMSEA = .02, WRMR = .84. As Figure 1 shows, the eight physical items had loadings on the physical construct ranging from .15 to .52 (only with loadings of V3 and V7 not statistically significant), while all 14 social items had statistically significant loadings on the social construct ranging from .25 to .57.

Given that the Y-ToM social and physical factors were strongly correlated with each other ($\phi = .95$, $p < .001$) in the two-factor model, we also tested it against an alternative model of a one-factor (unidimensional). The fit indices for the one-factor model were almost identical to the two-factor model, $\chi^2(209) = 223.47$, $p = .23$, CFI = .94, TLI = .94, RMSEA = .02, WRMR = .84. As a result, we proposed using the two-factor model as it was theoretically consistent with the structure of Happé's Strange Stories.

Convergent and Divergent Validity

The Y-ToM (social) and Strange Stories (social) latent factors were regressed onto age, sex, verbal IQ, and nonverbal IQ scores (see Figure 2). This model fit the data well, $\chi^2(0) = .00$, $p = .00$, CFI = 1.00, TLI = 1.00, RMSEA = .00, SRMR = .00, and revealed that the Y-ToM and Strange Stories (social) latent factors remained strongly correlated ($\phi = .94$, $p < .001$).

Subsequently, we examined correlates of individual differences in performance on the Y-ToM latent factor where the Y-ToM (social and physical) latent factors were regressed onto age, sex, verbal IQ, nonverbal IQ, and the 8 CBCL subscales (see Figure 3). This model provided an excellent fit to the data, $\chi^2(0) = .00$, $p = .00$, CFI = 1.00, TLI = 1.00, RMSEA = .00, SRMR = .00. Performance on the Y-ToM (social) was positively and significantly related to verbal IQ (.31). In addition, the Y-ToM (social) factor but not the Y-ToM (physical) factor was positively related to the CBCL Anxious (.24) and negatively related to the CBCL Social Problem (-.26) subscales scores.

Diagnostic Validity

We also compared the performance on the Y-ToM social total scores by using the CBCL Social Problem T-score cutoff of 60 ($N = 42$ who score 60 and above) against randomly selected cases in the sample who scored below 60 ($N = 42$) using a one-way ANOVA to examine diagnostic validity. There were no significant group differences in their verbal IQ ($p = .18$), age ($p = .58$) and sex ($p = .38$). Result indicated that there was a significant group difference on the Y-ToM social total score, $F(1, 82) = 8.51$, $p = .005$, $\eta 2 = .09$, ($M = 19.38$ vs. $M = 21.50$).

Hypothesis Testing on Gender as Part of Construct Validity

Females scored higher than males on the Y-ToM social total score, $F(1, 168) = 6.32$, $p = .01$, $\eta 2 = .36$, ($M = 21.23$ vs. $M = 19.88$) but there was no significant sex difference on the physical stories total score, $F(1, 168) = 2.20$, $p = .14$, $\eta 2 = .13$, ($M = 12.01$ vs. $M =$

11.51). Even after controlling for participants' verbal IQ and age, this pattern was still observed, $F(1, 166) = 8.18$, $p = .005$, $\eta 2 = .05$.

Evidence for Criterion Validity

Concurrent Validity

To examine the concurrent validity of the Y-ToM with the Strange Stories, a two latent factor measurement model was tested in which each item of the Y-ToM social items was loaded onto a single latent factor and each social item of the Strange Stories task was loaded onto a second correlated latent factor. Using the WLSMV estimator, results indicated a generally acceptable fit to the data, $\chi^2(251) = 283.71$, $p = .08$, CFI = .89, TLI = .88, RMSEA = .03, WRMR = .89. The standardised item loadings were all significant except for S18 (loaded onto Strange Stories) (see the upper panel of Figure 4). There was a strong correlation between the Y-ToM (social) and Strange Stories (social) latent factors ($\phi = .84$, $p < .001$).

We also examined a two latent factor measurement model in which each item of the Y-ToM physical items was loaded onto a single latent factor and each physical item of the Strange Stories task was loaded onto a second correlated latent factor. Using the WLSMV estimator, results indicated a generally acceptable fit to the data, $\chi^2(103) = 117.93$, $p = .15$, CFI = .90, TLI = .88, RMSEA = .03, WRMR = .82. The standardised item loadings were all significant except for V7 (loaded onto Y-ToM) and S4 (loaded onto Strange Stories) (see the lower panel of Figure 4). There was a strong correlation between the Y-ToM and Strange Stories (physical items) tasks ($\phi = .90$, $p < .001$).

Evidence for Reliability

The Y-ToM social subscale's internal consistency was .65 and did not improve meaningfully with the removal of any specific item. Spilt-half reliability based on the Spearman-Brown was .63. The Y-ToM physical subscale's internal consistency was lower

(.37) and did not improve meaningfully with the removal of any specific item. Spilt-half reliability based on the Spearman-Brown was .41.

Inter-rater reliability of the Y-ToM social and physical scores using a (two-way mixed) intra-class correlation and coefficient was .81, $p < .001$, 95% CI (.75, .86) and .89, $p < .001$, 95% CI (.85, .92) respectively. As for the test–retest reliability of the Y-ToM, the (two-way mixed) intra-class correlation coefficient was .57, $p < .001$, 95% CI (.33, .74) for physical scores and .45, $p = .001$, 95% CI (.19, .65) for social scores.

Discussion

This study presents the development and preliminary psychometric properties of a culturally adapted video version of Strange Stories test as a measure of advanced ToM for youths aged 13 to 16 years old (Y-ToM) in an Asian context (Singapore). Given that ToM has been implicated in the development of appropriate social communication and relationships, a suitable instrument is needed if not urgently and the Y-ToM seeks to provide a localised and easy-to-use tool for teachers or school professionals to identify significant deficits in ToM in youths.

There are several potential advantages of using a video version of the Strange Stories as compared to the traditional pencil-and-paper Strange Stories version. For instance, our video presentation of the social scenarios will appear to have higher face validity from the examinee's perspective than the pencil-and-paper (i.e., written presentation) equivalents. Another advantage of the video presentation is that the scenes are dynamic but yet the exact scene can be paused/stopped as a teaching moment for the students or client, which makes the Y-ToM suitable to be used as a teaching tool to explain the nuances of the social situations. Regular school support staff could use the Y-ToM to support students with social difficulties in their schools.

To answer research question 1 about content validity, a pilot test was conducted to ensure that the content and language used in the videos were suitable for the target sample (youths between 13 to 16 years old). In terms of evidence for construct validity (to answer research question 2), for structural validity, the CFA results suggest that a two-factor structure could be used for the Y-ToM which was consistent with the Happé's Strange Stories structure of social and physical items. However, the data also suggest that a one-factor model could work for the Y-ToM. It has been suggested that it may be more productive to think about the various ways in which ToM ability or inferring others' minds unfolds in the real-world (cf. Apperly, 2012), and understanding and inferring others' mental states may be a multidimensional process that interacts with other abilities (e.g., emotion recognition) (Warnell & Redcay, 2019).

As for convergent and divergent validity, the Y-ToM social factor is significantly correlated to the Strange Stories social factor, after controlling for age, sex, verbal IQ, and nonverbal IQ scores. As expected, performance on the Y-ToM social but not the physical factor is positively related to verbal IQ and negatively related to the CBCL Social Problem subscale score. It has been suggested that lower ToM ability was associated with higher levels of conduct problems, hyperactivity/inattention and emotional problems, in adolescents with and without ASD (Leno et al., 2020). However, one unexpected finding was the positive relationship between the Y-ToM social factor and the CBCL Anxious subscale score as typically poor ToM abilities are associated with high anxious or depressive symptoms (Lecce et al., 2019). A possible explanation could be that individuals who are generally more anxious are also more accurate in ToM reasoning tasks due to their heightened ability to attend to cues and accurately infer others' mental states (Zainal & Newman, 2018).

Diagnostic validity of the Y-ToM social scale is also established when we compare youths who scored highly (i.e., 1 standard deviation above the mean) on the CBCL Social

Problem subscale as compared to youths who scored lower. Therefore, there is the potential for the Y-ToM to be used to identify youths with possible social difficulties in community settings.

Lastly, there is a significant sex difference (small to medium effect size) where females performed better than the males on the Y-ToM social subscale, and even after accounting for their verbal IQ and age. This finding is consistent with other literature where it has been suggested that females were better in understanding social situations and intentions (Białecka-Pikul et al., 2017) and adds to the construct validity data of the relationships between ToM and gender (e.g., Calero et al., 2013; Shahrivar et al., 2017). In terms of concurrent validity (as part of criterion validity to answer research question 3), there are significant correlations among the Y-ToM social and physical scores with the Happé's Strange Stories social and physical scores.

As for evidence of reliability (to answer research question 4), the Y-ToM social subscale's internal consistency is considered acceptable in exploratory research (Nunally & Bernstein, 1994) though the Y-ToM physical scale's internal consistency is relatively poor (.37). It could be that the physical items were measuring different underlying factors even though they are supposedly to be a measure of general story or narrative comprehension (as opposed to mental-state reasoning). Since the content of the Y-ToM physical stories were similar to the Happé's original physical stories, it might be helpful to examine the internal consistency of the Happé's original physical stories. In this case, we are able to compute and the Cronbach's alpha is .52 while spilt-half reliability based on the Spearman-Brown is .45. It could be possible that the physical items in both the Y-ToM and Happé's stories do not reflect the underlying factor of narrative comprehension ability well in our Singapore sample of youths. In fact, little research has also been done to examine the factor structure of the

Happé's Strange Stories, with the exception of a study by Devine and Hughes (2013) looking at only four items of the social scale which loaded onto a single latent factor.

Using an intra-class correlation, inter-rater reliability of the Y-ToM social and physical scores is good (.81 to .89) while the test–retest reliability of the Y-ToM social and physical scores is lower (.45 to .57). Our students also perform better on the concept of white lie which could be due to Asian students seeing white lie as a pro-social behaviour as a way to consider other people's feelings (e.g., Cheung et al., 2015). This study extends our finding where Asians are better in detecting white lie to youths aged 13-16 years old as compared to the previous research which were done on children ranging from 7 to 11 years old (Cheung et al., 2015; Ma et al., 2011) and young adults aged 19 to 21 years old (Cheung et al., 2015).

Limitations and Future Directions

Several limitations of this study deserve attention. Due to a community sample of 170 youths in the current study, it would be pertinent to conduct a replication study with a larger and more diverse sample size to further examine the reliability and validity of the Y-ToM, and whether a one-factor or two-factor structure will be more appropriate. We also recognise that this study may only be suitable for measuring individual differences in ToM that are observed among typically developing youths and more work is needed to ascertain the utility for measuring clinically significant deficits in ToM in a clinical sample (e.g., youths with ASD). Some specific directions for future research could also be to explore the possibility of extending the Y-ToM to different age ranges such as the young adults (16 to 21 years old) and to examine possible factors that might inhibit or enhance such research.

Next, we acknowledge that some literature has highlighted flaws in the Happé's Strange Stories, including low internal consistency and poor task comprehension though the sample included younger children (aged 7 to 13 years old) (Hayward & Homer, 2017). In the current study, no general comprehension control questions of the Y-ToM or Strange Stories

were presented. However, we anticipate that the typically developing youths with adequate cognitive functioning should have no issues with comprehension though future studies may need to test out such comprehension concerns empirically. In addition, the mental state categories used in Y-ToM may not be exhaustive of the entire Happé's original Strange Stories (that include concepts such as pretend play and appearance/reality) but we are also cognisant of the time constraint placed on student participants to complete all the tests. As a result, we are not able to include all categories of ToM, and so we would like to propose using the seven mental state categories in Y-ToM as a starting point because these categories may be more relevant to our target age group.

Another possible limitation is the lack of time limit for Happe's Strange Stories task which might confound the validity analysis. However, there is no correlation between the time spent responding and the scores on Happe's Strange Stories task (p = .29 for physical items, p = .42 for social items). In addition, the fact that the ToM data come directly from the adolescents but the CBCL data come from parents might create a problem of accurately measuring the social dimension. It is also acknowledged that inter-rater reliability was only done for the items in Y-ToM and not for the Strange Stories.

Future research could further explore some other measurement properties within the COSMIN framework (Mokkink et al., 2010) that were not covered in the current investigation. For example, it may be conducive to examine the responsiveness of Y-ToM in detecting the effects of an intervention targeted at advanced ToM skills in youth. Kane (2013) proposed a four-stage framework to validity arguments, including *Scoring*, *Generalization*, *Extrapolation*, and *Implications* (also see details in Cook et al., 2015). Since the current investigation primarily focused on establishing the initial psychometric evidence for the Y-ToM, our work has mainly arrived at the stage of Scoring (i.e., translating the participants' responses on the Y-ToM into a score), and partly explored the stage of Generalization (i.e.,

using the score to reflect the participants' performance in another test [i.e., Happe's Strange Stories in this case]). Future research would be needed to extend this line of research to a higher level of validation, such as predicting the participants' performance in real-life social situations based on their Y-ToM scores (i.e., Extrapolation), and making decisions on whether intervention should be provided based on participants' Y-ToM scores (i.e., Implications).

Conclusion

In summary, this study describes the development of the Y-ToM based on Happé's Strange Stories test in video format and also to provide some preliminary psychometric properties of it. Preliminary results suggested that this tool, Y-ToM may be used as a valid and reliable instrument to measure advanced ToM in community samples of Singaporean youths.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth and Families.

Ang, R. P., Rescorla, L. A., Achenbach, T. M., Ooi, Y. P., Fung, D. S., & Woo, B. (2012). Examining the criterion validity of CBCL and TRF problem scales and items in a large Singapore sample. *Child Psychiatry & Human Development*, *43*(1), 70-86. doi:10.1007/s10578-011-0253-2

Apperly, I. (2010). *Mindreaders: The cognitive basis of" theory of mind"*. Psychology Press.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology, 65*(5), 825-839. https://doi.org/10.1080/17470218.2012.676055

Białecka-Pikul, M., Kołodziejczyk, A., & Bosacki, S. (2017). Advanced theory of mind in adolescence: Do age, gender and friendship style play a role?. *Journal of Adolescence*, *56*, 145-156. https://doi.org/10.1016/j.adolescence.2017.02.009

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241-251. https://doi.org/10.1017/S0021963001006643

Brewer, N., & Young, R. L. (2015). *Crime and autism spectrum disorder: Myths and mechanisms*. London: Jessica Kingsley.

Brewer, N., Young, R. L., & Barnett, E. (2017). Measuring Theory of Mind in adults with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *47*, 1927-1941. https://doi.org/10.1007/s10803-017-3080-x

Calero, C., Salles, A., Semelman, M., & Sigman, M. (2013). Age and gender dependent development of Theory of Mind in 6-to 8-year's old children. *Frontiers in Human Neuroscience, 7*, 281. https://doi.org/10.3389/fnhum.2013.00281

Cameron, C. A., Lau, C., Fu, G., & Lee, K. (2012). Development of children's moral evaluations of modesty and self-promotion in diverse cultural settings. *Journal of Moral Education, 41*(1), 61-78. https://doi.org/10.1080/03057240.2011.617414

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

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. https://doi.org/10.1111/medu.12678

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

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., ... & Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders, 36*(5), 623-636. https://doi.org/10.1007/s10803-006-0107-0

Happé, 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*(2), 129-154. https://doi.org/10.1007/BF02172093

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*(3), 454-462. https://doi.org/10.1111/bjdp.12186

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73. https://doi.org/10.1111/jedm.12000

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Lecce, S., Ceccato, I., & Cavallini, E. (2019). Theory of mind, mental state talk and social relationships in aging: The case of friendship. *Aging & Mental Health, 23*(9), 1105-1112. https://doi.org/10.1080/23311908.2018.1487270

Leno, V. C., Chandler, S., White, P., Yorke, I., Charman, T., Jones, C. R. G., Happe, F., Baird, G., Pickles, A., & Simonoff, E. (2020). Associations between theory of mind and conduct problems in autistic and nonautistic youth. *Autism Research*, 14(2), 276-288. https://doi.org/10.1002/aur.2346

Livingston, L. A., Carr, B., & Shah, P. (2019). Recent advances and new directions in measuring theory of mind in autistic adults. *Journal of Autism and Developmental Disorders, 49*(4), 1738-1744. https://doi.org/10.1007/s10803-018-3823-3

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

McCrimmon, A. W., & Smith, A. D. (2013). Review of the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II). *Journal of Psychoeducational Assessment, 31*, 337–341. https://doi.org/10.1177/0734282912467756

McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology press.

Meinhardt-Injac, B., Daum, M. M., & Meinhardt, G. (2020). Theory of mind development from adolescence to adulthood: Testing the two-component model. *British Journal of Developmental Psychology*, *38*(2), 289-303. https://doi.org/10.1111/bjdp.12320

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*, 737-745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Murray, K., Johnston, K., Cunnane, H., Kerr, C., Spain, D., Gillan, N., Hammond, N., Murphy, D., & Happé, F. (2017). A new test of advanced theory of mind: The "Strange Stories Film Task" captures social processing differences in adults with autism spectrum disorders. *Autism Research, 10*(6), 1120-1132. https://doi.org/10.1002/aur.1744

Muthén, B. O., & Muthén, L. K. (2017). *Mplus (Version 8)*.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

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

Ooi, Y. P., Rescorla, L., Ang, R. P., Woo, B., & Fung, D. S. (2011). Identification of autism spectrum disorders using the child behavior checklist in Singapore. *Journal of Autism and Developmental Disorders*, *41*(9), 1147-1156. doi:10.1007/s10803-010-1015-x

Perez-Zapata, D., Slaughter, V., & Henry, J. D. (2016). Cultural effects on

mindreading. *Cognition*, *146*, 410-414. https://doi.org/10.1016/j.cognition.2015.10.018

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515-526. https://doi.org/10.1017/S0140525X00076512

Rescorla, L., Achenbach, T., Ivanova, M. Y., Dumenci, L., Almqvist, F., Bilenberg, N., Bird, H., Chen, W., Dobrean, A., Dopfner, M., Erol, N., Fombonne, E., Fonseca, A., Frigerio, A., Gritens, H., Hannesdottir, H., Kanbayashi, Y., Lambert, M., Larsson, BO., . . . & Verhulst, F. (2007). Behavioral and emotional problems reported by parents of children ages 6 to 16 in 31 societies. *Journal of Emotional and Behavioral Disorders*, *15*(3), 130-142. https://doi.org/10.1177/10634266070150030101

Salisch, M., Zeman, J., Luepschen, N., & Kanevski, R. (2014). Prospective relations between adolescents' social-emotional competencies and their friendships. *Social Development*, *23*(4), 684-701. https://doi.org/10.1111/sode.12064

Shahrivar, Z., Tehrani-Doost, M., Khorrami Banaraki, A., Mohammadzadeh, A., & Happé, F. (2017). Normative data and psychometric properties of a Farsi translation of the strange stories test. *Autism Research, 10*(12), 1960-1967. https://doi.org/10.1002/aur.1844

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

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. (2011). *Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II)*.

San Antonio, TX: Pearson.

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

Wolters, N., Knoors, H., Cillessen, A. H., & Verhoeven, L. (2014). Behavioral, personality, and communicative predictors of acceptance and popularity in early adolescence. *The Journal of Early Adolescence*, *34*(5), 585-605. https://doi.org/10.1177/0272431613510403

Zainal, N. H., & Newman, M. G. (2018). Worry amplifies theory-of-mind reasoning for negatively valenced social stimuli in generalized anxiety disorder. *Journal of Affective Disorders, 227*, 824-833. https://doi.org/10.1016/j.jad.2017.11.084

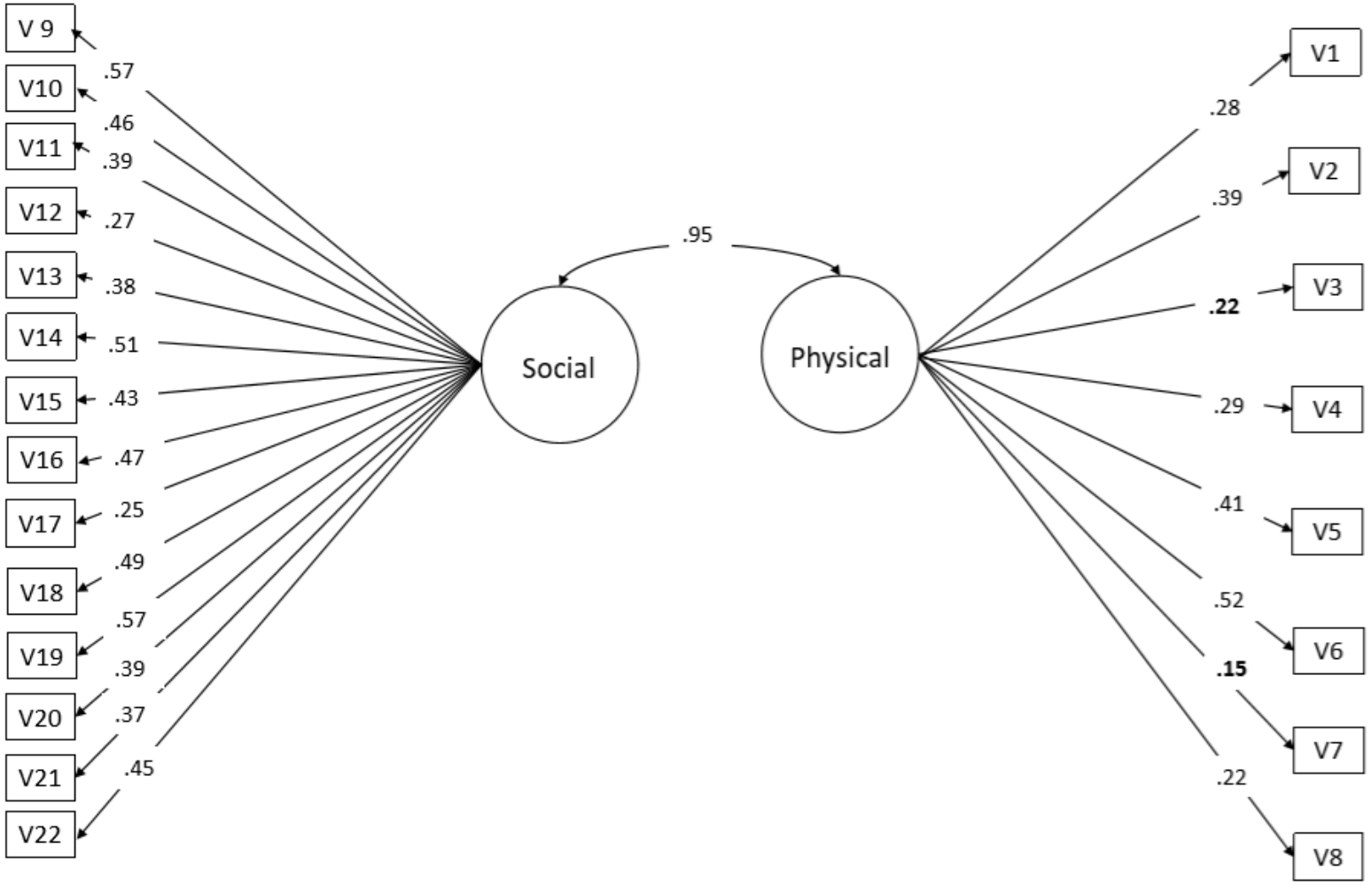

Figure 1. Diagram of CFA results for Y-ToM. Note. Loadings on V3 and V7 are bold to indicate non statistical significance.

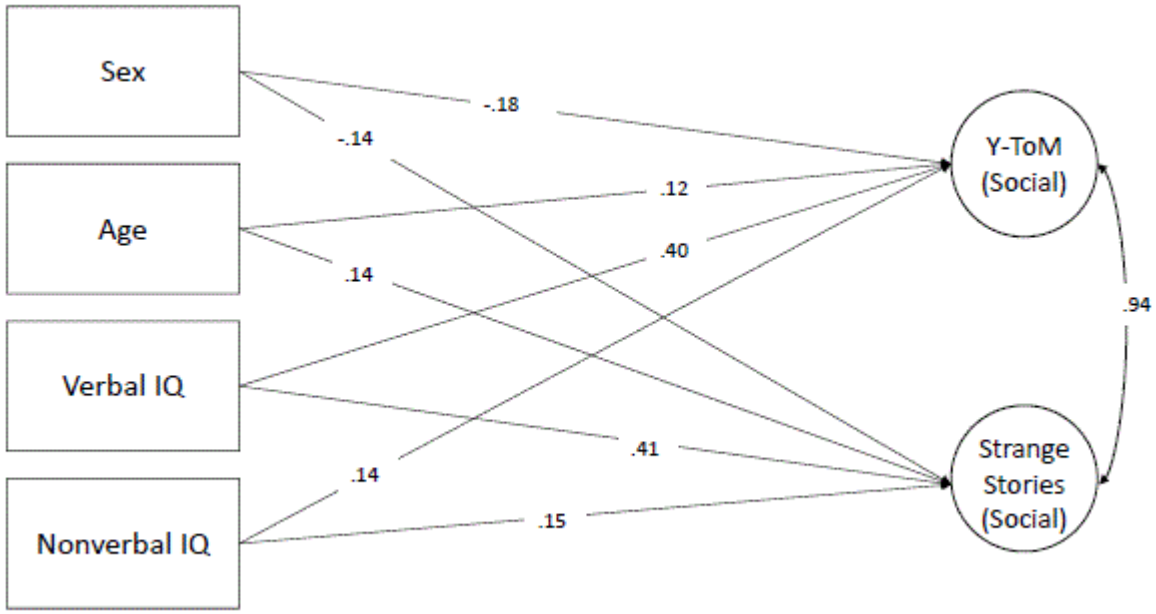

Figure 2. Diagram of Latent Modelling for Correlates of Y-ToM (Social) and Strange Stories (Social) Factors.

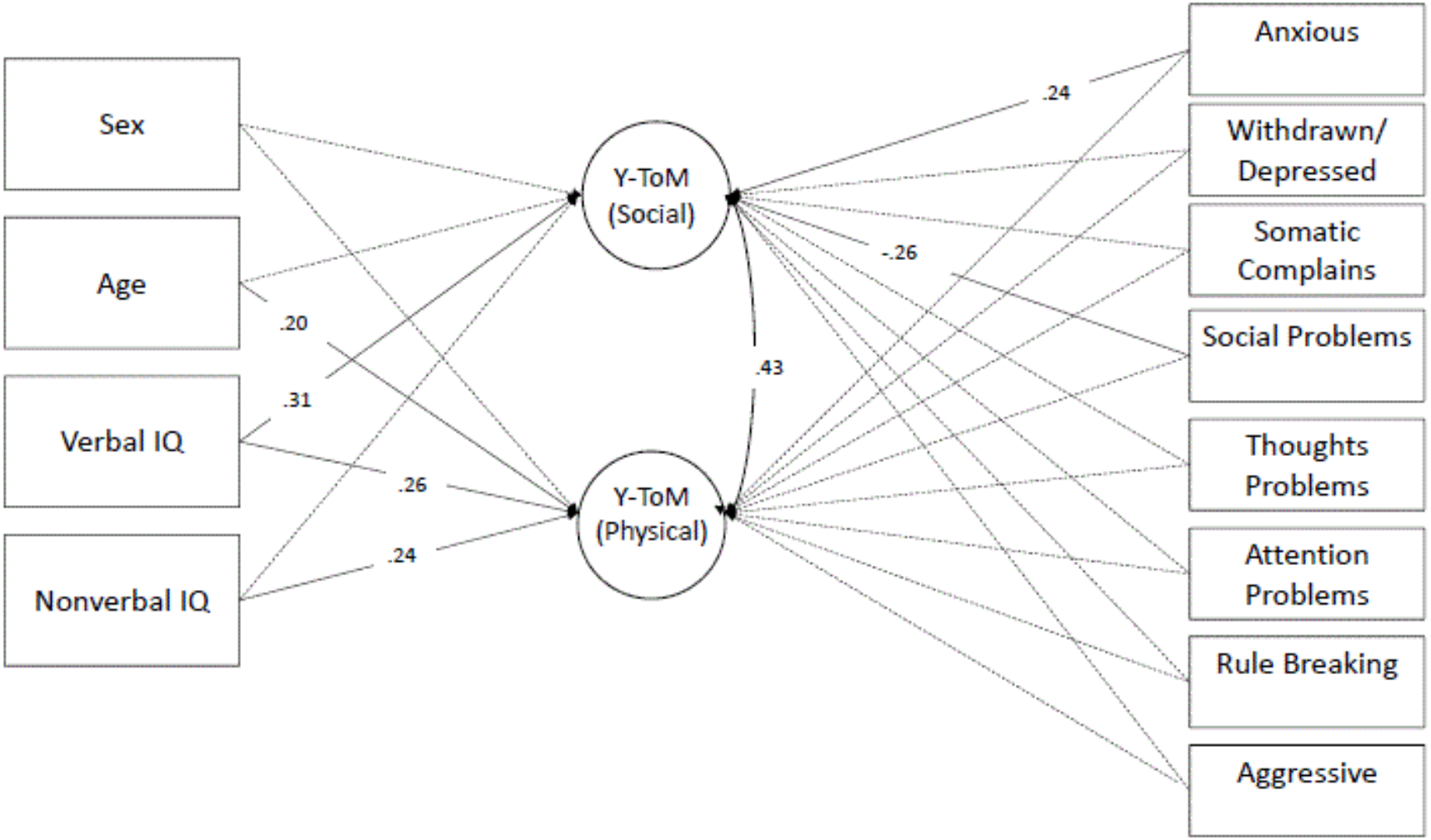

Figure 3. Diagram of Latent Modelling for Correlates of Y-ToM and Strange Stories Factors with CBCL Variables. Dashed paths are non-significant.

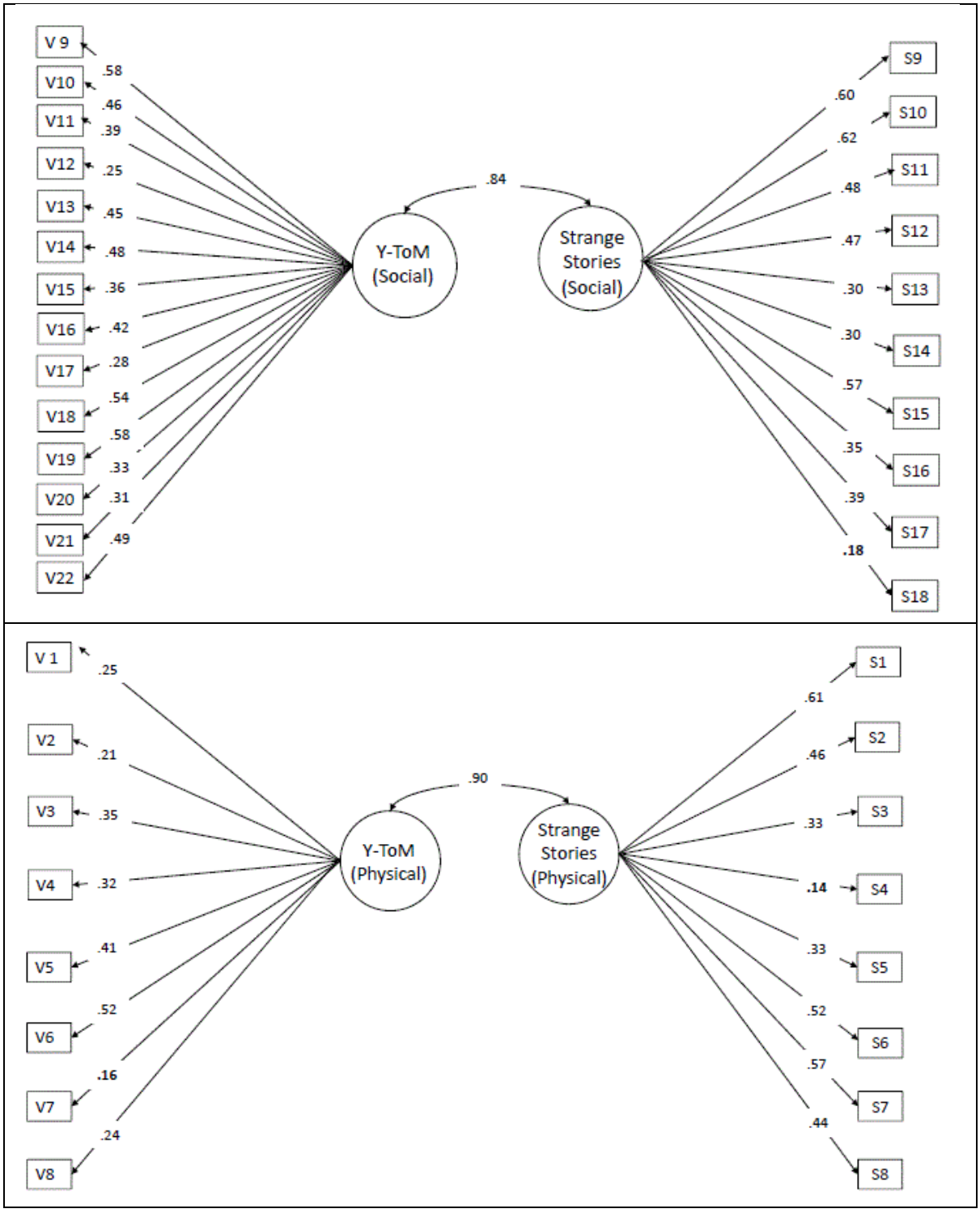

Figure 4. Diagram of Latent Modelling for Concurrent Validity of Y-ToM and Strange Stories' Social and Physical Items.

Appendix A

Example of a Mental State story script (https://www.youtube.com/watch?v=yivzR4uDx_E)

White Lie (Present)

Rachel is celebrating her birthday with Joan and Kate.

Kate: Since it's your 21st birthday special, the two of us decided to get you a gift.

Joan passes Rachel a present.

Kate: I'm sure you will love it.

Rachel: Thank you. What is it? Oh is it the speaker I've always been looking at?

(opening the present) (opens to reveal a night light)

Kate: mmm.. not exactly.

Rachel: Oh.. it's a … (disappointed) night light?

Joan: Yeah. Do you like it?

Rachel: (quick change) Yeah yea. I love it! I didn't expect it at all. Thank you so much.

Question: Why did Rachel say that "she loves the gift"?

Example of a Physical State story script (https://www.youtube.com/watch?v=MdylINXa2mI)

Paula is selling donuts and walks over.

Paula : Hi, I'm selling some snacks. One for just one dollar and 5 for four dollars. Would you like to get some?

Aaron: Actually I only need one, but sure, give me 5.

Question: Why did he buy 5 donuts?