Measuring Theory of Mind in Middle Childhood: Reliability and Validity of the Silent Films and Strange Stories Tasks

Rory T. Devine

Claire Hughes

University of Cambridge

Author Note

R.T. Devine and C. Hughes, University of Cambridge, Center for Family Research, Department of Psychology, Free School Lane, Cambridge, CB23RQ, United Kingdom. We would like to thank all the principals, teachers and pupils at the participating schools in Cambridge, Nottingham and London, England. We would also like to thank Abby Furniss and Sakshi Rathi for their assistance with data collection and coding. R.T. Devine was funded by the Isaac Newton Trust, Cambridge. For more information about using the Silent Film Task in research contact R.T. Devine by electronic mail to: rtd24@cam.ac.uk.

Key Words: *Theory of Mind, Middle Childhood, Adolescence, Reliability, Validity.*

Measuring Theory of Mind in Middle Childhood: Reliability and Validity of the Silent Films and Strange Stories Tasks

Abstract

Recent years have seen a growth of research on the development of children's ability to reason about others' mental states (or 'theory of mind') beyond the narrow confines of the preschool period. The overall aim of this study was to investigate the psychometric properties of a task battery comprised of items from Happé's (1994) Strange Stories task and Devine and Hughes' (2013) Silent Film task. 460 ethnically and socially diverse children (211 boys) aged between 7 and 13 years old completed the task battery at two time points separated by 1 month. The Strange Stories and Silent Film tasks were strongly correlated even when verbal ability and narrative comprehension were taken into account and all items loaded onto a single theory-of-mind latent factor. The theory-of-mind latent factor provided reliable estimates of performance across a wide range of theory of mind ability and showed no evidence of differential item functioning across gender, ethnicity or SES. The theory-of-mind latent factor also exhibited strong one-month test retest reliability and this stability did not vary as a function of child characteristics. Taken together these findings provide evidence for the validity and reliability of the Strange Stories and Silent Film task battery as a measure of individual differences in theory of mind suitable for use in middle childhood. We consider the methodological and conceptual implications of these findings for research on theory of mind beyond the preschool years.

Key Words: *Theory of Mind, Middle Childhood, Adolescence, Reliability, Validity, Psychometrics, Measurement*

Measuring Theory of Mind in Middle Childhood: Reliability and Validity of the Silent Films and Strange Stories Tasks

How children learn to use mental states, such as desires, knowledge and beliefs to predict and explain others' behavior (commonly referred to as the acquisition of a 'theory of mind') is a topic that has attracted extensive theorizing and empirical research for almost four decades (for recent reviews, see Hughes & Devine, 2015; Wellman, 2014). Most of this research has centered on a single task, the false belief task (Wimmer & Perner, 1983), in which an object is moved in an agent's absence, such that children need to recognise that the agent has a mistaken belief in order to predict or explain his or her behavior. More complex tasks measure children's ability to attribute beliefs to an agent about another agent's beliefs (i.e., 'second-order' false beliefs) (Perner & Wimmer, 1985) or to attribute emotional states to others on the basis of false beliefs (e.g., Harris et al., 1989). These tasks have been used to study both individual differences and age-related changes in theory of mind (ToM) in the preschool and early school years (e.g., Wellman, Cross & Watson, 2001).

In the past decade the developmental scope of ToM research has been greatly increased by the design of new tasks for use with infants (e.g., Luo & Baillargeon, 2007) and with adults (e.g., Apperly, Samson & Humphreys, 2009). With some notable exceptions including early research on the development of children's understanding of the interpretive nature of knowledge in early middle childhood (e.g., Carpendale & Chandler, 1996) and evidence for meaningful individual differences in pre-adolescents' ability to reason about characters' mental states (e.g., Bosacki & Astington, 1999), the developmental period of middle childhood has been largely overlooked. However, in recent years this developmental period has begun to attract research attention (e.g., Apperly et al., 2011; Banerjee, Watling & Caputi, 2011; Devine & Hughes, 2013; Dumontheil et al., 2010).

Middle childhood (the developmental period between ages 6 and 12) is a developmentally interesting period in which to study ToM. From a socio-cultural perspective, it is worth noting that in primary school children are exposed to increasingly more sophisticated forms of knowledge (e.g., fictional literature) and also spend increasing amounts of time outside the home interacting with their peers in a greater variety of contexts (e.g., Del Giudice, 2014; Eccles, 1999). Understanding how these new experiences shape and are shaped by individual differences in ToM presents a novel opportunity for researchers. Indeed recent work in the field has demonstrated that individual differences in ToM in this period are related to important social and academic outcomes (e.g., Banerjee, Watling & Caputi, 2011; Lecce, Caputi & Hughes, 2011). From a neuropsychological perspective, there is evidence of continued structural changes in the frontal and parietal lobes (specifically, grey matter volume increases in these regions across middle childhood) (e.g., Giedd et al., 1999) and related gains in cognitive performance in domains such as executive function across middle childhood (e.g., Davidson, Amso, Anderson & Diamond, 2006). Research on ToM in middle childhood could shed light on the correlates and consequences of these neuro-psychological changes. Indeed researchers have now begun to examine the developmental links between ToM and executive function in middle childhood in order to understand the factors underpinning the continued development of ToM in this period (e.g., Bock, Galllaway & Hund, 2014; Lagattuta, Sayfan & Blattman, 2010; Lagattuta, Sayfan & Harvey, 2014). In an effort to contribute to this budding new field of research, the focus of the current study was to examine the validity and reliability of two tasks that appear promising as developmentally appropriate and useful indicators of ToM in middle childhood: the Strange Stories Task (Happé, 1994) and a more recent analogue task using brief clips from a classic silent film (Devine & Hughes, 2013).

Validity and Reliability of the False Belief Task

Validity refers to whether or not a test is measuring the construct that it purports to measure (e.g., Rust & Golombok, 2008). Test validity is established through the accumulation of evidence about whether the test conforms to expectations and hypotheses about the construct being measured (Carmine & Zellar, 1979; Rust & Golombok, 2009). Tasks that purport to measure a particular construct (e.g., false belief understanding) should be related to tasks that measure the same or similar constructs (convergent validity) and unrelated to tasks that do not (discriminant validity). Ideally tests should also show evidence of correlations with real-life outcomes (criterion validity). Test validity can be established by examining the correlations between concurrent measures and longitudinal outcomes and by assessing group differences (e.g., 3-year-old children versus 4-year-old children, typical versus atypical groups) (Cronbach & Meehl, 1955; Messick, 1995).

Four sources of evidence support the validity of the false belief task. Firstly, children's performance on the different versions of the false-belief task show moderate to strong concurrent correlations and so appear to measure a single construct (e.g., Hughes et al., 2014; Hughes et al., 2000). That is, various false-belief tasks show convergent validity. Secondly, there is now a growing body of evidence supporting the criterion validity of false belief tasks: individual differences in performance on the false-belief task among typically developing children can be predicted by early social experiences (e.g., parent-child talk about mental states) and in turn correlate with important social outcomes (Hughes & Devine, 2015*;* Slaughter, Imuta, Peterson & Henry, 2015). Thirdly, the false belief task is sensitive to development: between the ages of 2 and 5 children's performance on this task improves dramatically (Wellman, Cross & Watson, 2001). Finally, children with known impairments in social competence (e.g., children with ASD and 'Hard to Manage' preschoolers) show marked deficits in performance on the false belief task relative to children matched in age and verbal ability (e.g., Baron-Cohen, Leslie & Frith, 1985; Hughes et al., 1998). In sum, there is

evidence that the false belief task shows convergent, discriminant and criterion validity as a measure of ToM.

Test validity hinges upon the precision and repeatability of a measurement, that is, the reliability of the test (e.g. Carmines & Zellar, 1979). Test reliability can be established through examining the dimensionality of a set of items that comprise a test (i.e., the internal consistency of items within a test) and through examining the stability of test scores across time (i.e., the test-retest reliability of a measure) (e.g., Rust & Golombok, 2009). These two forms of reliability testing enable researchers to establish the precision with which individual test items measure the construct of interest and the extent to which test scores can be reproduced with repeated measurement. Test reliability is vital to the study of individual differences and developmental change. If error variance is not accounted for, it is difficult to know whether observed correlations or test score changes reflect genuine associations with or changes in the construct of interest. Evidence for the reliability of the false belief task has grown in the past two decades. In an early study, Hughes et al. (2000) demonstrated that a battery of first- and second-order false-belief tasks exhibited good internal consistency and strong one-month test-retest reliability. Importantly, by examining the interaction between initial task performance and individual differences in verbal ability, Hughes et al. (2000) found that the test-retest reliability of the task battery was stable across different levels of verbal ability suggesting that the battery of tasks could be used reliably with children of varying levels of ability.

In recent years researchers have begun to apply modern psychometric approaches when assessing the validity and reliability of tests of ToM for young children (e.g., Wellman & Liu, 2004). Confirmatory factor analysis (CFA) is a flexible way in which to assess the psychometric properties of test batteries. CFA is hypothesis driven and permits researchers to test a measurement model against data using multiple fit indices. CFA enables researchers to

tackle the task impurity problem inherent in cognitive research by partitioning the variance that is common between a set of items or tasks (i.e., the true score variance) from the variance associated with a specific task or item and measurement error (i.e., the residual variance) (e.g., Miyake et al., 2000). Importantly for test development, CFA enables researchers to examine the stability of a measurement model (or 'measurement invariance') across different groups (e.g., gender, ethnic groups) and over time (Brown, 2006). Establishing measurement invariance is an important step in studying the fairness of a test. Differences in test performance should reflect genuine differences in the variable of interest and not group differences in the psychometric properties of the test (Knight & Zerr, 2010; Millsap, 2010). Differential item functioning (DIF) occurs when groups differ in their performance on a particular item because that item involves abilities other than those the item was intended to measure and the groups differ on those abilities rather than the target ability and so can undermine the fairness of a test (Walker, 2011). Using multiple groups CFA it is possible to assess whether items exhibit DIF.

CFA has been applied extensively in the study of executive function and has been used to analyse the psychometric properties of test batteries designed to measure EF in early childhood. For example, Willoughby and colleagues have used CFA and Item Response Theory (IRT) to examine the measurement structure, precision and test-retest reliability of a novel battery of EF tasks for young children (e.g., Willoughby, Blair, Wirth & Greenberg, 2010; Willoughby & Blair, 2011). These studies demonstrate the flexibility of using CFA as a means to assess the psychometric properties of task batteries and researchers have begun to apply CFA to assess individual differences in performance on measures of false belief understanding (Hughes, Ensor & Marks, 2011; Hughes et al., 2014). These findings have revealed that false-belief task batteries load onto a single latent factor and are invariant across gender and partially invariant across cultures. In summary, the false belief task appears to

provide a valid and reliable measure of ToM for young children. A parallel body of evidence that assesses the psychometric properties of ToM tasks for use in middle childhood is now needed to support research on individual differences and change in ToM in this developmental period.

Validity and Reliability of 'Advanced' Measures of Theory of Mind

In the past decade researchers have devised a diverse range of tasks purported to measure different aspects of ToM use with a variety of stimuli, such as: vignettes, cartoons, audio recordings and film clips. In addition to the wide range of stimuli employed by researchers, the tests appear to measure distinct aspects of ToM use such as: emotion understanding, perspective taking, understanding the interpretive nature of mind, attribution of intention and explanation of behavior with reference to beliefs, knowledge and desires (e.g., Baron-Cohen et al., 1997; Carpendale & Chandler, 1996; Castelli et al., 2000; Dumontheil et al., 2010; Dziobek et al., 2006; Golan et al., 2006; Happé, 1994). Supporting the validity of these tasks, adults with ASD and schizophrenia have been shown to have difficulties on these tasks relative to matched 'neurotypical' controls (Chung, Birch & Strube, 2014). Crucially these limitations in performance are specific to test items centered on mentalistic content and not simply on narrative understanding or non-mental content (e.g., White et al., 2009). While there is some evidence for the validity of these 'advanced' tasks, less is known about the precision and stability of these measures. With few exceptions little effort has been made to evaluate the psychometric properties of these tasks (e.g. Dziobek et al., 2006; Fernandez-Abascal et al., 2013).

In an effort to develop age-appropriate ToM tasks to study individual differences and age-related changes in ToM in middle childhood, Devine and Hughes (2013) administered Happé's (1994) vignette-based Strange Stories task alongside a novel Silent Film task to 230 middle-class children aged between 8 and 13. Successful performance on both of these tasks

required children to explain a character's behavior with reference to the character's knowledge, beliefs and desires. The findings from this initial study revealed that both the Strange Stories and Silent Film tasks were sensitive to age-related differences in performance with neither task exhibiting marked ceiling effects. There were strong concurrent associations between the two tasks supporting the convergent validity of the Silent Film task with the widely-used Strange Stories task. More recently longitudinal findings have shown that performance on a battery of false-belief tasks at age 6 was significantly correlated with later performance on both the Strange Stories and Silent Film tasks at age 10 (Devine, White, Ensor & Hughes, *submitted*). These findings provide further evidence for the convergent validity of these advanced ToM tasks. Supporting the criterion validity of these tasks, low scores in girls were associated with self-reported loneliness and low scores in boys were associated with self-reported peer exclusion. Despite differences in the modality of each task, CFA revealed that a unidimensional ToM latent factor underpinned performance on the diverse items of the Strange Stories and Silent Film tasks. This ToM latent factor exhibited measurement invariance in boys and girls with no evidence of differential item functioning. In sum, the Strange Stories and Silent Film task battery is a promising way to measure ToM in middle childhood. That said further work is needed to investigate the validity, precision and reliability of these tasks. The purpose of our study was to investigate the psychometric properties of the Strange Stories and Silent Film task battery in a large ethnically and socially diverse sample of children aged between 7 and 13 years.

Our first aim was to examine further the concurrent, discriminant and construct validity of the Strange Stories and Silent Film tasks. While it is tempting to claim that a ToM latent factor underpins participants' performance on the items of the Strange Stories and Silent Film tasks the relations between the items may simply reflect common variance due to another variable, for example, the ability to comprehend a narrative sequence rather than

mental states *per se*. To rule out this alternative interpretation, the participants completed three 'control' stories (matched in length and linguistic complexity) that described scenarios involving human characters but contained no mental-state content (White et al., 2009) to determine whether the correlation between performance on the Strange Stories (mental state items) and Silent Film task items persisted once individual differences in story or narrative comprehension were taken into account.

The second aim of our study was to examine the precision and measurement invariance of the Strange Stories and Silent Film task battery. Using Item Response Theory (IRT) models it was possible to compute standard errors that are conditional on a certain trait or 'theta' level and so assess the reliability of the Strange Stories and Silent Film task battery at different levels of latent ability (Embretson & Reise, 2000; Hays et al., 2000). Extending findings about the measurement invariance of the Strange Stories and Silent Film task battery across boys and girls, the diverse sample recruited for the current study also made it possible to assess measurement invariance across different ethnic and socio-economic groups.

The third aim of our study was to investigate the test-retest reliability of the Strange Stories and Silent Film task battery. To date, no published studies have sought to examine the short-term stability of measures of ToM in middle childhood and adolescence. Latent variable modelling with CFA provides a particularly robust way in which to examine test-retest reliability. Typically, researchers estimate the correlation between initial and retest scores. Since this approach does not account for item-specific variance and measurement error the correlations between test scores might not provide accurate estimates of the stability of performance on the latent variable. In one pioneering study, Willoughby and Blair (2011) examined the one-month test-retest reliability of a battery of executive function tasks for preschool children using a latent variable approach. By accounting for the potential instability of item-specific variance, Willoughby and Blair (2011) found that in contrast to the moderate

test-retest correlations between specific items, the correlation between the latent factors approached unity. We adopted the same analytic strategy the current study to examine the one-month test-retest reliability of the Strange Stories and Silent Film tasks. Given the large and diverse sample, we also assessed whether test-retest stability was moderated by child characteristics such as age, gender, ethnicity, socio-economic status and verbal ability. Building on the analysis used by Hughes et al. (2000) when studying the test-retest stability of the false belief task battery, non-significant interaction effects between child characteristics and test stability would provide evidence to support the applicability of the Strange Stories and Silent Film task battery across a diverse range of children.

To summarize, our study had three primary aims. Our first aim was to examine the convergent, discriminant and construct validity of the Silent Film and Strange Stories tasks as measures of ToM suitable for use across middle childhood. Our second aim was to assess the precision and measurement invariance of the Silent Film and Strange Stories tasks. Our third aim was to assess the test-retest reliability of the Silent Film and Strange Stories tasks and examine whether one-month test-retest stability varied as a function of individual differences child characteristics.

Method

Participants

Participants were recruited from 8 socio-economically and ethnically diverse state schools in the South East of England. The eight schools involved in this study were average or above average in terms of total number of pupils (i.e., >263 pupils for primary schools and > 978 pupils for secondary schools) and all were based in urban areas (OFSTED, 2014). Of the 565 children in the classes approached, 38 children were not eligible to take part because teachers reported that the children had developmental disabilities or spoke English as an additional language. Of the remaining 527 children, 460 children (87%) completed both testing

sessions. Those children who did not complete both sessions either opted out of the study or were not in attendance at one of the sessions. These 460 children (69.9% White British) included 249 girls and 211 boys aged between 7.32 and 13.34 years (at the initial visit). There were 45 children aged 7 (24 females, *M* Age = 7.66, *SD* = .20), 110 children aged 8 (66 females, *M* Age = 8.50, *SD* = .28), 61 children aged 9 (27 females, *M* Age = 9.47, *SD* = .31), 82 children aged 10 (52 females, *M* Age = 10.48, *SD* = .26), 51 children aged 11 (27 females, *M* Age = 11.34, *SD* = .23), 86 children aged 12 (40 females, *M* Age = 12.35, *SD* = .30) and 25 children aged 13 (13 females, *M* Age = 13.07, *SD* = .08). At the time of data collection, children whose parents were in receipt of state income support in the UK were entitled to free school meals. Twenty percent of participants were recruited from schools that were above average in terms of the number of children receiving free school meals (> 26%). This school-level data about socio-economic status gleaned from official statistics (OFSTED, 2014) provided an objective indicator of material deprivation at school.

Procedure

To maximize the number of participants included in our study, we used a passive consent (‘opt out’) procedure. This procedure was approved by the University Psychology Research Ethics Committee and is consistent with national research ethics guidelines (British Psychological Society, 2010). Specifically, we first sought permission from the head teacher and classroom teacher to conduct the study. Then, approximately one week prior to data collection, we sent an information letter about the study to parents and guardians explaining the procedures and purpose. Parents and guardians were requested to contact the school teacher or research team if they wished to ‘opt out’ of the study. In total 2 parents declined consent for their children to participate in the study. At the start of each session the researchers explained the procedures of the study to the participants. Children were advised that they did not have to return their response booklets at the end of the session if they did not

wish to participate. Those children who did not wish to take part were given alternative activities to complete by their class teachers. Teachers and classroom assistants were present during all testing sessions. At the end of the second session, we provided the participants with a verbal debrief and explained the purpose of the research, how the research will be analysed and answered any additional questions the children had.

The children completed two 50-minute researcher-led whole-class testing sessions approximately 1 month apart, M = 31.40 days, SD = 6.25, Range: 21 – 41 days. During the first session, the children completed a short demographic questionnaire, a verbal ability test and the two ToM tasks. The order of the Strange Stories and Silent Film tasks were counter-balanced across classes and separated by the verbal ability test. In the second session, the children completed a demographic questionnaire and the ToM tasks. Given that testing took place during whole-class sessions, we took a number of steps to ensure the validity of results. First, throughout each session the children were instructed to complete each task in silence and not to share their answers. Second, there were two researchers and (at least) two members of teaching staff present at each testing session. This meant that the children could, if needed, receive appropriate support.

Measures

Strange Stories Task. The Strange Stories task (Happé, 1994) consisted of five mental state stories depicting social situations, each followed by a single open-response question that required participants to explain a character's behavior with reference to his/her mental states. The children also administered three control stories (matched in length and linguistic complexity) that described scenarios involving human characters but contained no mental-state content (White et al., 2009). These three items were included to provide a measure of general story or narrative comprehension (as opposed to mental-state reasoning). Each control story was followed by a question which required the children to explain an

event in the story. Figure 1 (Panel A) shows an example vignette from the mental state and control stories alongside the scoring rubric. The text of each vignette appeared on an over-head projector. The researcher then read the text aloud to the class. The text and the question were left on the over-head screen until all children had written a response.

Participants' responses were scored using the coding scheme refined by White et al. (2009). For the mental state stories, correct responses that involved explicit mentalizing scored 2 points; partially correct responses that fell short of a full explanation scored 1 point; and inaccurate or irrelevant responses scored 0 points. For the control stories, correct responses received 2 points; partially correct responses received 1 point and incorrect or inaccurate responses received 0 points. Individual mental state stories exhibited moderate to strong inter-rater reliability, Mean $\kappa = .82$, Range: $.79 \leq \kappa \leq .85$, all ps $< .01$, as did individual control stories, Mean $\kappa = .84$, Range: $.74 \leq \kappa \leq .95$, all ps $< .01$.

Silent Film Task. The Silent Film Task (Devine & Hughes, 2013) consisted of 5 short film clips from a classic silent comedy depicting instances of deception, misunderstanding and false belief. Figure 1 (Panel B) depicts the events from a sample clip in which a van driver accidentally locks Harold (the main character) in his van and the coding scheme for responses to that item. The participants watched each clip once and after each clip, the researcher read the question aloud to the class. The researcher did not play the next clip until all children had written a response. Participants' responses were scored using a rating scheme developed by Devine and Hughes (2013): full understanding (2 points) was awarded if a participant provided an accurate mentalistic explanation; partial understanding (1 point) was awarded if the participant provided a correct response that fell short of a mentalistic explanation; participants failed (0 points) an item if the response was irrelevant or factually inaccurate. Individual items exhibited moderate to strong inter-rater reliability, Mean $\kappa = .81$, Range: $.76 \leq \kappa \leq .93$, all ps $< .01$.

Mill Hill Vocabulary Scale. Participants completed the multiple-choice section of the Mill Hill Vocabulary Scale (MHVS) (Rust, 2008) to measure verbal ability. The MHVS was designed to index receptive vocabulary in 7 to 18 year olds in group settings. In each item, the children were asked to select a synonym for a target word from 6 possible response options. Children were awarded 1 point for each correctly identified synonym. The number of correct items were summed together to give a total raw score (possible range: 0 – 44).

Results

Analytic Strategy

The data were analysed using a latent variable framework in *Mplus* Version 7 (Muthèn & Muthèn, 2012). Given the categorical nature of our data we used a mean- and variance-adjusted weighted least squares estimator (rather than a maximum likelihood estimator) in each of our models (Brown, 2006; Kline, 2011). For each model we evaluated fit using Brown's (2006) four recommended criteria: a non-significant χ^2 test; Comparative Fit Index (CFI) ≥ .90; Tucker Lewis Index ≥ .90; Root Mean Square Error of Approximation (RMSEA) ≤ 0.08.

Descriptive Statistics

Table 1 shows the mean scores on each of the key variables at the initial and retest visit. Verbal ability scores were normally distributed and ranged from 3 to 31. Boys and girls were matched in age, M_{Boys} = 10.30 years, SD_{Boys} = 1.75, M_{Girls} = 10.08, SD_{Girls} = 1.69, t (458) = -1.40, p = .16, and in verbal ability, M_{Boys} = 15.88, SD_{Boys} = 5.33, M_{Girls} = 15.08, SD_{Girls} = 5.11, t (458) = -1.63, p = .10. Table 1 also shows the proportion of children who failed, received partial credit and passed each item of the Silent Film Task and Strange Stories (Mental State) Task. Inspection of summed scores at the Initial Visit revealed that only 3.9% of children performed at ceiling on the Strange Stories (Mental State) Task and 0.7% of children performed at ceiling on the Silent Film Task. Moreover, the distributions for both

tasks were symmetrical and did not significantly deviate from normality. The total score for the three items of the Strange Stories (Control) Task ranged from 0 to 6 points and was normally distributed. There were moderate correlations between performance on each of the items of the Strange Stories and Silent Film tasks across the test-retest interval (Table 1). These test-retest correlations did not differ by gender, SES or age.

Validity of the Strange Stories and Silent Film Tasks

Our first aim was to examine the convergent validity of the Silent Film and Strange Stories (Mental State) tasks as measures of individual differences in ToM. A two latent factor measurement model in which each item of the Silent Film Task loaded onto a single latent factor and each mental-state item of the Strange Stories (Mental State) Task loaded onto a second correlated latent factor provided an excellent fit to the data, χ^2 (43) = 40.99, p = .56, RMSEA = 0.00, CFI = 0.99, TLI = 0.99. The standardized item loadings were all significant (see Figure 2). There was a strong correlation between the Silent Film and Strange Stories (Mental State) tasks, ϕ = .85, p < .001.

Next the Silent Film and the Strange Stories (Mental State)latent factors were regressed onto age, gender, ethnicity, SES and verbal ability (See Figure 3, Panel A). This model fit the data well, χ^2 (88) = 100.47, p = .17, RMSEA = 0.02, CFI = 0.97, TLI = 0.96, and revealed that the Silent Film and Strange Stories (Mental State) latent factors remained strongly correlated, ϕ = .73, p < .001. Given the possibility that the overlap between the two latent factors could be explained by narrative comprehension, we tested a second model in which each latent factor was regressed onto a latent factor representing performance on the Strange Stories Control Items, verbal ability, age, gender and SES. This model provided a good fit to the data, χ^2 (118) =141.40, p = .07, RMSEA = 0.02, CFI = 0.98, TLI = 0.98. Performance on the Silent Film and Strange Stories (Mental State) latent factors remained strongly correlated, ϕ = .66, p < .001. To examine this further a model in which the Strange

Stories Control latent factor and Silent Film latent factor were regressed onto the Strange Stories (Mental State) latent factor, verbal ability, age, gender, ethnicity and SES. This model showed that there was no significant correlation between children's performance on the Silent Film latent factor and the Strange Stories Control latent factor when performance on the Strange Stories (Mental State) latent factor, verbal ability, age, gender, ethnicity and SES were taken into account, $\phi = -.14$, $p = .73$.

Given the strong overlap between the Silent Film and Strange Stories (Mental State) latent factors, a single latent factor measurement model in which the items from both tasks loaded onto a single ToM latent factor was assessed. The one factor model provided an excellent fit to the data (see Figure 2 for standardized parameter estimates), χ^2 (44) =45.11, p = .42, RMSEA = 0.01, CFI = 0.99, TLI = 0.99. To examine correlates of individual differences in performance on the ToM latent factor, the ToM latent factor was regressed onto age, gender, ethnicity, SES and verbal ability (see Figure 3, Panel B). This model accounted for 51% of the variance in the ToM latent factor and provided a good fit to the data, χ^2 (94) = 115.21, p = .07, RMSEA = 0.02, CFI = 0.95, TLI = 0.94. Consistent with previous findings, performance on the ToM latent factor increased with age (independently of verbal ability). Girls out-performed boys (despite being matched in age and verbal ability). Ethnicity was unrelated to performance on the ToM latent factor but children from less affluent schools lagged behind their more affluent peers in terms of ToM performance.

Together the findings from these three models provide evidence for the convergent validity of the Silent Film task with the Mental State Items from the Strange Stories task and suggest that both tasks measure a single latent ability. Importantly, the findings also provide new evidence for the discriminant validity of these tasks: the overlap between the Silent Film and Strange Stories (Mental State) tasks could not be accounted for by narrative comprehension (as measured by the Control Items of the Strange Stories Task) or verbal

ability. The construct validity was further strengthened by evidence that the ToM latent factor showed expected correlations with age, verbal ability and gender.

Precision and Invariance of the Strange Stories and Silent Film Task Battery

Our second aim was to examine the measurement precision and invariance of the Silent Film and Strange Stories ToM latent factor. First, given that the individual items of both tests consisted of ordered categories, we used a graded item response theory (IRT) model using robust maximum likelihood estimation to assess the precision of the ToM battery at different levels of the latent ToM factor (Embretson & Reise, 2000; Muthèn & Muthèn, 2012). Findings from the total information curve revealed that the ToM task battery was most precise when testing participants performing between 2SD below the mean and 1SD above the mean (see Figure 4). For participants with average levels of ToM ability (i.e., where $\theta = 0$), the reliability co-efficient was .72. The reliability co-efficient for those participants performing at -2SD and +1SD was .68. The task battery provided less precise estimates for those participants performing at 3SD above the mean (.39) but adequate estimates for those performing at 3SD below the mean (.63).

Next, we used multiple groups CFA to test the measurement invariance of the ToM latent factor across gender, ethnicity and SES using a series of nested multiple groups CFAs to examine the measurement invariance of the ToM latent factor in boys and girls, white and non-white children and high and low socio-economic groups. In each case the first (baseline) model tested the assumption of equal form or factor structure in both groups and the second (invariance) model tested for differential item functioning (DIF) by assessing the fit of a model with equal form, equal factor loadings and equal item thresholds in both groups. Changes in model fit (as measured by a corrected χ^2 difference test suitable for use with WLSMV estimation – Muthèn & Muthèn, 2012) and inspection of modification indices were used to assess for the presence of DIF. The results of these nested models are presented in

Table 2. In sum, there was no evidence for the presence of DIF in the items making up the ToM latent factor suggesting that items were equally fair across gender, ethnicity and SES.

Test-Retest Reliability of the Strange Stories and Silent Film Task Battery

Our third and final aim was to examine the test-retest reliability of the Strange Stories and Silent Films task battery. To assess the test-retest reliability of the ToM latent factor, we specified a two latent factor model in which items from the Strange Stories and Silent Film tasks the initial visit loaded onto one latent factor and corresponding items from the retest visit loaded onto a second correlated latent factor. To account for item-specific variance, the residual terms from each item at the initial visit was correlated with its corresponding item at the retest visit. This model provided a good fit to the data, χ^2 (197) = 247.13, RMSEA = 0.02, CFI = 0.98, TLI = 0.98. To test the metric invariance of this model the item loadings and latent factor variances were constrained to be equal across both time points. This model continued to provide a good fit to the data, χ^2 (208) =246.51, RMSEA = 0.02, CFI = 0.99, TLI = 0.98. The ToM latent factor exhibited metric invariance across time and had excellent test-retest reliability, ϕ = .83, p <.001 (see Table 3 for parameter estimates).

The next step was to determine whether the test-retest reliability of the Strange Stories and Silent Film task battery varied as a function of different child characteristics. Given that children's performance on both the Strange Stories and Silent Film tasks was correlated with age, SES, ethnicity, gender and verbal ability, we examined whether the test-retest reliability of the ToM latent factor varied as a function of these dimensions of individual differences. We specified a structural equation model in which we regressed the ToM latent factor from the retest visit onto the ToM latent factor from the initial visit, age, gender, ethnicity, SES and verbal ability. In addition to these variables we examined the multiplicative interaction between the ToM latent factor scores at the initial visit and age, gender, SES, ethnicity, and verbal ability. If any multiplicative interaction term was significant it would indicate that the

test-retest correlation varied across different groups of children undermining the utility of the test battery for that group. The non-standardized parameter estimates for this model are presented in Table 4. Together the model accounted for 51% of the variance in the ToM latent factor scores at the retest visit. Importantly, none of the multiplicative interaction terms were statistically significant. In summary, the ToM latent factor exhibited equal levels of test-retest reliability across gender, age and different levels of SES and verbal ability.

Discussion

This investigation of the psychometric properties of the Strange Stories and Silent Film task battery involved 460 children aged between 7 and 13 years and yielded three sets of findings. First, scores on both tasks were strongly correlated, even when verbal ability and narrative comprehension were taken into account. Replicating previous findings (Devine & Hughes, 2013), the ToM latent factor was sensitive to effects of age and gender. Second the Strange Stories and Silent Film task battery provided precise estimates of performance across a wide range of ToM ability, with no evidence of differential item functioning across gender, ethnicity or SES. Third, the ToM latent factor showed excellent one-month test-retest reliability, which did not vary as a function of different child characteristics. We will now discuss the implications of each of these findings for research on ToM in middle childhood.

Validity of the Strange Stories and Silent Film Task Battery

Our results provide new evidence about the discriminant validity of the task battery. Crucially, individual differences in narrative comprehension or in verbal ability did not explain associations between performance on items from the Strange Stories and Silent Film tasks. We also replicated and extended previous findings concerning the convergent validity for the Strange Stories and Silent Film Tasks and the construct validity of the combined task battery (Devine &Hughes, 2013). Specifically, we replicated the previously observed correlations between performance on the ToM latent factor, age and verbal ability and also

confirmed a small but significant advantage in ToM performance for girls compared with boys. We extended earlier work by recruiting a more ethnically and socially diverse sample of children which permitted us to examine the relations between SES, ethnicity and ToM performance. Consistent with findings from the preschool years (e.g., Cutting & Dunn, 1999), we found that SES was independently correlated with performance on the ToM latent factor in middle childhood: children from affluent schools outperformed their less affluent peers.

Additional support for the validity of this Strange Stories and Silent Film task battery as a measure of individual differences in ToM in middle childhood come from reports that performance on this task battery shows: (i) longitudinal associations with prior (age 6) performance on a battery false-belief tasks (Devine et al., *submitted*); and (ii) cross-sectional associations with self-reported peer acceptance (Devine & Hughes, 2013). Much like the false belief task, the extant data suggest support the convergent, discriminant and criterion validity of the Strange Stories and Silent Film task battery. Future work assessing group differences in performance on this task battery in children with known performance deficits in social understanding (e.g., children with ASD) and further evidence of links between task performance and other social outcomes will strengthen the validity of this task battery.

Methodological Implications

Our results regarding the precision, measurement invariance and longitudinal stability of performance on the Strange Stories and Silent Film task battery all provide evidence for the reliability of this task battery. These results have three important methodological implications. First, we used a whole-class testing approach to collect data. Although this approach could potentially have a number of drawbacks in terms of data quality (e.g., conferring between participants, potential distractions), it enabled us to recruit a large and diverse sample in a relatively short time. Indeed the sample size of the current study is far greater than that typically reported in the ToM literature. In addition, our findings indicate

that this approach to data collection yields reliable data, as measured by test precision, measurement invariance and test stability. This approach could greatly aid future work research on ToM in middle childhood and adolescence providing ready access to large datasets with sufficient power to examine the sometimes small to medium strength associations between individual differences in ToM and cognitive and social correlates (Devine & Hughes, 2014; Slaughter et al., 2015).

Second, our findings suggest that the items of the Strange Stories and Silent Film task battery provided precise estimates of ToM ability for children with average and below average levels of latent ability (with less precise estimates of performance at the upper end of ability). Moreover the task items provide unbiased estimates of ToM performance. That is, there was no evidence of differential item functioning that favored children of a particular gender, ethnic or socio-economic group. In terms of future work, this suggests that the Strange Stories and Silent Film task can be used across diverse groups of children. Indeed, the task battery has now been used to measure individual differences in ToM in children from Hong Kong (Wang et al., submitted) and Italy (Lecce et al., in preparation). Given that poor ToM performance tends to be related to negative social and relationship outcomes (e.g., Slaughter et al., 2015), having precise measurements of ToM at the lower end of the spectrum of performance is important. The addition of more challenging test items could enhance the reliability of the Strange Stories and Silent Film task battery further to permit research on the correlates and consequences being a 'virtuoso' mind-reader.

Third our results indicate that the Strange Stories and Silent Film task battery exhibited strong one-month test-retest reliability and that the stability of task performance was unrelated to a range of child characteristics. Methodologically, this is an important finding because it supports the application of the Strange Stories and Silent Film task battery in future longitudinal and intervention studies designed to learn about the developmental

individual differences in ToM in middle childhood. Indeed future work designed to explain ToM development beyond the preschool years hinges on longitudinal and intervention research. Given that the vast majority of cognitive and social accounts of ToM development are based on data from preschool children, longitudinal and intervention designs are needed to investigate whether ToM continues to exhibit developmental links with individual differences in language, executive function and social adjustment (Hughes & Devine, 2015).

Conceptual Implications

Apart from demonstrating that the Strange Stories and Silent Film task battery provides a reliable and valid measure of individual differences in ToM, our findings have broader implications for how age-related changes in ToM beyond the preschool years should be conceptualized. Our findings add to a small but growing body of research that indicates that ToM performance continues to improve with age across middle childhood and adolescence (Apperly et al., 2011; Baron-Cohen et al., 1999; Banerjee et al., 2011; Dumontheil et al., 2010). Interestingly, the effects of age on ToM performance were independent of verbal ability. Age-related changes in ToM performance beyond the preschool years can be interpreted in at least two ways (Miller, 2009). The first of these is the conceptual change account, in which it is posited that there are further conceptual breakthroughs and discoveries about the mind to be made beyond an understanding of desires, knowledge and beliefs (Flavell, 2004; Sullivan et al., 1994). When considered in the context of existing evidence about age-related changes in performance on ToM tasks, our findings challenge this conceptual change account. Our test items involved reasoning about desires, knowledge and beliefs but showed marked age-related changes in performance. Whether due to accumulating social experience or developments in domain-general cognitive abilities across middle childhood, children showed gradual gains in their ability to use their ToM in a diverse range of scenarios. This pattern of results is more consistent with a second

account, in which it is argued that children do not require further conceptual insights to succeed on 'advanced' ToM tasks but instead the complex requirements of these tasks challenge the correct application of children's insights about the mind (Apperly, 2012). From this perspective, children become more proficient at using their ToM appropriately. Longitudinal designs are needed to untangle the factors that contribute to developmental changes in ToM use across middle childhood.

Group differences in ToM performance between boys and girls and between affluent and less affluent children also deserve note. First, with regard to gender differences, our findings provide further evidence that, in middle childhood, girls outperform boys in tasks designed to measure mental-state reasoning (Bosacki & Astington, 1999; Calero, Salles, Semelman & Sigman, 2013; Devine & Hughes, 2013). While gender differences in mental-state reasoning are the basis of a central claim of Baron-Cohen's (2002) 'empathising-systemising' account of autism, such findings stand in contrast to the mixed evidence for gender differences in preschoolers' ToM (e.g. ; Hughes et al., 2011). These findings raise important questions about when gender differences in mentalising might emerge in the course of development and what factors might contribute to the female advantage in mental-state reasoning. One possibility is that different social experiences (e.g., gender contrasts in patterns of play) during middle childhood might give rise to differences in cognitive performance (Maccoby, 1966). Indeed current data indicates that girls typically outperform boys academically in middle childhood and adolescence (e.g., Deary, Strand, Smith & Fernandes, 2007). One interesting avenue for future study might be to address the educational and clinical implications of these gender differences in ToM. For example, recent research has indicated that children's sensitivity to teacher criticism mediates the developmental association between ToM task performance and later academic achievement in middle childhood (Lecce et al., 2011). Studies involving larger samples could be used to examine

whether gender differences in ToM (and sensitivity to teacher criticism) might partially explain the gender gap in academic performance in middle childhood.

Second, with regard to socio-economic differences in ToM performance, our findings extend existing data on the links between SES and ToM in the preschool years. In a recent meta-analysis of the existing evidence Devine and Hughes (*submitted*) reported a small but significant association between SES and false-belief understanding. The nature of the links between SES and ToM are poorly understood. According to the social selection account, individual cognitive characteristics can lead to lower levels of attainment, education and occupational prestige (Conger & Donnellan, 2007). The gap in cognitive performance between children from lower and higher socio-economic groups in domains such as ToM might therefore reflect the heritability of ToM. However, heritability estimates for individual differences in ToM are modest (e.g., Hughes et al., 2005), such that the social selection hypothesis is unlikely to account fully for the links between family SES and children's ToM. According to the social causation account, the limited resources associated with lower SES might hinder children's social and cognitive development (Bradley & Corwyn, 2002). By this account, factors such as parent-child interactions are likely to mediate the link between SES and children's ToM. Recent longitudinal evidence linking parental mental-state talk in the preschool years and ToM in middle childhood suggests that parent-child interactions might account for the continued effects of SES on ToM in middle childhood (Ensor et al., 2014).

Caveats and Conclusions

Before concluding, two potential limitations of our study should be mentioned. First, while whole-class testing permitted us to recruit a large and diverse sample of children, our procedures meant that we could not obtain individual level socio-economic data from parents. Instead we used school-level data rather than individual-level data as an indicator of SES. Further research incorporating more detailed measures of SES is needed to confirm our

findings. Second, although our sample was ethnically diverse we had insufficient numbers of children within each of the non-White categories to perform more detailed analyses of the measurement invariance of the task battery across different ethnic groups. Related to these points, the findings reported here may be specific to the particular cultural context of the UK and will need to be investigated further in different cultural settings.

Notwithstanding these potential limitations, our study marks an important contribution to the field both methodologically and conceptually. From a methodological point of view, the Strange Stories and Silent Film task battery provides a valid measure of individual differences in ToM. Moreover this task battery exhibits precision and measurement invariance across a wide range of children of different backgrounds as well as stability over time. The task battery is easy to administer and code and suitable for group-based testing with children between the ages of 7 and 13 which permits large-scale data collection for research on ToM in middle childhood. From a conceptual perspective, our findings provide further evidence for age-related gains and gender differences and new evidence about socio-economic differences in in ToM performance in middle childhood. Continued research on ToM in middle childhood and beyond will provide an exciting opportunity to investigate the causes, correlates and consequences of individual differences and age-related developments in mind-reading.

References

Apperly, I. A. (2012). What is theory of mind? Concepts, cognitive processes and individual differences. *Quarterly Journal Of Experimental Psychology*, *65*, 825-839.

Apperly, I. A., Samson, D., & Humphreys, G. W. (2009). Studies of adults can inform accounts of theory of mind development. *Developmental Psychology, 45,* 190 -201.

Apperly, I. A, Warren, F., Andrews, B. J., Grant, J., & Todd, S. (2011). Developmental continuity in theory of mind: Speed and accuracy of belief – desire reasoning in children and adults. *Child Development, 82,* 1691-1703.

Banerjee, R., Watling, D., & Caputi, M. (2011). Peer relations and understanding of faux pas: Longitudinal evidence for bidirectional associations. *Child Development, 82,* 1887 - 1905.

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

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with Autism or Asperger Syndrome. *Journal of Child Psychology and Psychiatry, 38,* 813-822.

Bock, A.M., Gallaway, K.C., & Hund, A.M. (2014). Specifying the links between executive functioning and theory of mind during middle childhood: Cognitive flexibility predicts social understanding. *Journal of Cognition and Development*. doi: 10.1080/15248372.2014.888350.

Bosacki, S. & Astington, J.W. (1999). Theory of mind in preadolescence: Relations between social understanding and social competence. *Social Development, 8,* 237 – 255.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53,* 371-99.

British Psychological Society (2010). *Code of Human Research Ethics*. Leicester, UK: British Psychological Society.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. London: The Guilford Press.

Calero, C., Salles, A., Semelman, M. & Sigman, M. (2013). Age and gender dependent development of theory of mind in 6 to 8 year old children. *Frontiers in Human Neuroscience, 7*, Article 281, 1 – 6.

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

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage, 12*, 314-25.

Chung, Y. S., Barch, D., & Strube, M. (2014). A meta-analysis of mentalizing impairments in adults with schizophrenia and autism spectrum disorder. *Schizophrenia Bulletin*, *40*, 602-616.

Conger, R. D., & Donnellan, M. B. (2007). An interactionist perspective on the socioeconomic context of human development. *Annual Review of Psychology*, *58*, 175-199.

Cronbach, L. J., & Meehl, P. E. (1995). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281 - 302.

Cutting, A. L., & Dunn, J. (1999). Theory of mind, emotion understanding, language and family background: individual differences and interrelations. *Child Development*, *70*, 853 - 865.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition and task switching. *Neuropsychologia*, *44*, 2037 - 2078.

Deary, I.J., Strand, S., Smith, P. & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13 – 21.

Del Giudice, M. (2014). Middle childhood: An evolutionary-developmental synthesis. *Child Development Perspectives, 8*, 193 - 200.

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.

Devine, R. T., & Hughes, C. (2014). Relations between false-belief understanding and executive function in early childhood: A meta-Analysis. *Child Development, 85*, 1777 - 1794.

Devine, R.T. & Hughes, C. (*in preparation*). Family correlates of false-belief understanding: A meta-analytic review. Unpublished Manuscript.

Devine, R. T., White, N., Ensor, R., & Hughes, C. (*submitted*). Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. Unpublished Manuscript.

Dumontheil, I., Apperly, I. A., & Blakemore, S.J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science, 13*, 331 - 338.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., et al. (2006). Introducing MASC: A movie assessment of social cognition. *Journal of Autism and Developmental Disorders, 36*, 623 - 636.

Ensor, R., Devine, R. T., Marks, A., & Hughes, C. (2014). Mothers' cognitive references to 2-year-olds predict theory of mind at ages 6 and 10. *Child Development, 85*, 1222-1235.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fernandez-Abascal, E. G., Cabello, R., Fernandez-Berrocal, P., & Baron-Cohen, S. (2014). Test-retest reliability of the "reading the mind in the eyes" test: a one-year follow-up study. *Molecular Autism, 4*, 33 - 38.

Flavell, J. H. (2004). Theory-of-mind development: retrospect and prospect. *Merrill-Palmer Quarterly, 50*, 274 - 290.

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., et al. (1999). Brain development during childhood and adolescence: a longitudinal MRI study. *Nature Neuroscience, 2*, 861-863.

Golan, O., Baron-Cohen, S., & Hill, J. (2006). The Cambridge Mindreading Face-Voice Battery: Testing complex emotion recognition in adults with and without Asperger syndrome. *Journal of Autism and Developmental Disorders, 36*, 169 - 183.

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

Harris, P. L., Johnson, C. N., Hutton, D., Andrews, G., & Cooke, T. (1989). Young children's theory of mind and emotion. *Cognition & Emotion, 3*, 379 - 400.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*, 28 - 42.

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test-retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *Journal of Child Psychology and Psychiatry, 41*, 483 - 490.

Hughes, C., Dunn, J., & White, A. (1998). Trick or treat? Uneven understanding of mind and emotion and executive dysfunction in "hard-to-manage" preschoolers. *Journal of child psychology and psychiatry, 39*, 981-994.

Hughes, C., & Devine, R. T. (2015). A social perspective on theory of mind. In M. E. Lamb (Ed.), *Handbook of Child Psychology and Developmental Science (Volume III): Socioemotional Processes* (pp. 564 - 609). Hoboken, NJ: Wiley.

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,* Article ID: 893492.

Hughes, C., Ensor, R., & Marks, A. (2011). Individual differences in false belief understanding are stable from 3 to 6 years of age and predict children's mental state talk with school friends. *Journal of Experimental Child Psychology*, *108,* 96-112.

Hughes, C., Jaffee, S., Happé, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child Development*, *76*, 356 - 370.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd ed.). London: Guilford Press.

Knight, G. P., & Zerr, A. A. (2010). Measurement equivalence in child development research. *Child Development Perspectives*, *4*, 1-4.

Lagattuta, K.H., Sayfan, L., & Blattman, A.J. (2010). Forgetting common ground: Six- to seven-year-olds have an over-interpretive theory of mind. *Developmental Psychology*, *4*6, 1417 – 1432.

Lagattuta, K.H., Sayfan, L. & Harvey, C. (2014). Beliefs about thought probability: Evidence for persistent errors in mindreading and links to executive control. *Child Development, 85,* 659 – 674.

Lecce, S., Bianco, F., Devine, R. T., & Hughes, C. (*in preparation*). Longitudinal relations between theory of mind and executive function in middle childhood. Unpublished Manuscript.

Lecce, S., Caputi, M. & Hughes, C. (2011). Does sensitivity to criticism mediate the relationship between theory of mind and academic achievement? *Journal of Experimental Child Psychology, 110,* 313 – 331.

Luo, Y., & Baillargeon, R. (2010). Toward a mentalistic account of early Psychological Reasoning. *Current Directions in Psychological Science, 19,* 301-307.

Maccoby, E. E. (1966). Sex differences in intellectual functioning. In E. E. Maccoby (Ed.), *The Development of Sex Differences* (pp. 25 - 55). Stanford, CA: Stanford University Press.

Messick, S. (1995). Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Millsap, R. E. (2010). Testing Measurement Invariance Using Item Response Theory in Longitudinal Data: An Introduction. *Child Development Perspectives, 4,* 5-9.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive Psychology, 41,* 49-100.

Muthen, L. K., & Muthen, B. O. (2012). *Mplus: Statistical Analysis With Latent Variables. User's Guide.* (7th ed.). Los Angeles, CA.: Muthen and Muthen.

OFSTED (2014). *Office for Standards in Education, Children's Services and Skills (OFSTED) School Data Dashboard*. Retrieved from http://www.dashboard.ofsted.gov.uk

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that..." Attribution of second-order beliefs by 5-to 10-year old children. *Journal of Experimental Child Psychology, 39,* 437 - 471

Rust, J. (2008). *Raven's Standard Progressive Matrices and Mill Hill Vocabulary Scale*. London: Pearson Education

Rust, J., & Golombok, S. (2009). *Modern Psychometrics: The Science of Psychological Assessment.* (3rd ed.). London: Routledge.

Slaughter, V., Imuta, K., Peterson, C., & Henry, J. D. (2015). Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development*. doi: 10.1111/cdev.12372.

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology, 30,* 395 - 402.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment, 29*(4), 364 - 376.

Wang, Z., Devine, R. T., Wong, K. K. Y., & Hughes, C. (*submitted*). Theory of mind and executive function in middle childhood across cultures. Unpublished Manuscript.

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. Oxford, UK: Oxford University Press.

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.

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

White, S., Hill, E., Happe, F., & Frith, U. (2009). Revisiting the Strange Stories: Revealing mentalising impairments in Autism. *Child Development, 80,* 1097 - 1117.

Willoughby, M., & Blair, C. (2011). Test-retest reliability of a new executive function battery for use in early childhood. *Child Neuropsychology, 17,* 564-579.

Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2010). The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. *Psychological Assessment, 22,* 306 - 317.

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

Table 1. *Across Time Performance on Individual Items of the Silent Film Task and Strange Stories Task*

		Initial Visit		% Participants			Retest Visit		% Participants			Spearman's *r*
		M	*SD*	Fail	Partial	Pass	*M*	*SD*	Fail	Partial	Pass	.39
SF	Why do the men hide?	0.30	0.53	73.9	22.7	3.5	0.36	0.55	67.2	29.5	3.3	.48
SF	What is the woman thinking?	1.33	0.87	26.3	14.1	59.6	1.43	0.81	20.0	16.8	63.2	.37
SF	Why does the driver lock Harold in the van?	1.35	0.89	28.5	7.6	63.9	1.46	0.85	23.3	7.4	69.9	.38
SF	What is the delivery man feeling and why?	1.40	0.67	10.7	38.5	50.9	1.45	0.66	9.3	36.5	54.1	.34
SF	Why did Harold pick up the cat?	1.11	0.93	38.6	11.8	49.7	1.28	0.87	27.6	17.2	55.2	.32
SF	Why did Harold fan Mildred?	1.09	0.94	40.4	10.1	49.5	1.15	0.91	34.6	16.0	49.3	.40
SS	Brian's Story	1.14	0.49	5.6	74.3	19.8	1.15	0.42	2.6	79.9	17.5	.40
SS	Mrs Peabody's Story	1.53	0.59	4.8	37.8	57.4	1.62	0.56	3.7	30.3	65.9	.52
SS	The Prisoner's Story	1.20	0.91	33.1	13.4	53.5	1.35	0.84	24.0	17.0	59.0	.42
SS	Simon's Story	1.34	0.64	9.4	47.6	43.0	1.37	0.62	7.2	48.6	44.2	.48
SS	The Burglar's Story	0.87	0.81	39.6	33.6	26.8	0.97	0.72	27.7	47.7	24.6	.39
	Silent Film Task Summed Total Score	6.57	2.47	-	-	-	7.11	2.39	-	-	-	-
	Strange Stories Task (Mental) Total Score	6.05	2.22	-	-	-	6.43	2.04	-	-	-	-
	Strange Stories Task (Control) Total Score	3.83	1.44	-	-	-	-	-	-	-	-	-
	Mill Hill Vocabulary Scale Total Score	15.45	5.22	-	-	-	-	-	-	-	-	-

Note. SF = Silent Film Task. SS = Strange Stories Task.

Table 2. *Comparison of Nested Models for Testing Measurement Invariance of the Theory of Mind Latent Factor.*

Model	χ^2 (*df*)	***CFI***	***TLI***	***RMSEA***	$\Delta\chi^2$
Gender					
Baseline Model	105.88 (98)	0.99	0.99	0.02	-
Invariance Model	109.01 (108)	0.99	0.99	0.01	6.48
Ethnicity					
Baseline Model	100.55 (98)	0.99	0.99	0.01	-
Invariance Model	106.88 (108)	1.00	1.00	0.01	8.02
Socio-Economic Status					
Baseline Model	122.36 (98)	0.94	0.93	0.03	-
Invariance Model	126.88 (108)	0.95	0.94	0.03	7.40

Note. Baseline Models = Equal Form/Factor Structure. Invariance Model = Equal Form, Equal Factor Loadings and Equal Item Thresholds.

Table 3. *Standardised WLSMV Estimates* for *Theory of Mind Latent Factor Loadings at Time 1 and Time 2 and Residual Covariances from Time 1 to Time 2*

	Theory of Mind Latent Factor Loading		Residual
	Time 1	Time 2	Covariance
SF Why do the men hide?	.30	.31	.50
SF What is the woman thinking?	.15*	.13*	.66
SF Why does the driver lock Harold in the van?	.39	.25	.56
SF What is the delivery man feeling and why?	.49	.48	.45
SF Why did Harold pick up the cat?	.30	.25	.51
SF Why did Harold fan Mildred?	.46	.53	.37
SS Brian's Story	.59	.60	.37
SS Mrs Peabody's Story	.57	.62	.36
SS The Prisoner's Story	.63	.64	.65
SS Simon's Story	.51	.55	.44
SS The Burglar's Story	.60	.55	.54

Note. *$p < .05$. All other loadings and covariances were significant, $p < .01$. SF = Silent Film Task. SS = Strange Stories Task.

Table 4. *Test-Retest Reliability Moderator Analyses Unstandardized Estimates*

Predictor	**Unstd. Est.**	***S.E.***	***Z***	***P***
ToM at Initial Visit	1.26	0.43	2.97	.003
Age (Years)	0.13	0.03	3.95	.001
Gender	-0.24	0.09	-2.82	.005
Socio-economic Status	-0.52	0.13	-4.09	.001
Verbal Ability	0.05	0.01	4.19	.001
Ethnicity	-0.21	0.09	-2.18	0.03
ToM x Age	-0.05	0.04	-1.24	.22
ToM x Gender	0.02	0.11	0.14	.89
ToM x Socio-economic Status	-0.21	0.16	-1.37	.17
ToM x Ethnicity	0.05	0.11	0.37	.71
ToM x Verbal Ability	-0.03	0.02	-1.57	.12

Note. Dependent Variable = Theory of Mind Latent Factor Scores at Retest Visit. ToM = Theory of Mind Latent Factor Scores.

Figure 1. Sample Items and Coding Schemes from the Strange Stories and Silent Film Task Battery.

Panel A. Sample Strange Stories Items

Strange Stories: Mental State Story (White et al., 2009)

Simon is a big liar. Simon's brother Jim knows this. He knows that Simon never tells the truth! Now yesterday Simon stole Jim's table-tennis paddle, and Jim knows Simon has hidden it somewhere, though he can't find it. He's very cross. So he finds Simon and he says, 'Where is my table-tennis paddle? 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 paddle is under his bed.

Why will Jim look in the cupboard for the panel?
2 – Reference to Jim knowing Simon lies
1 – Reference to the facts (that's where it really is, Simon is a big liar) or Simon hiding it without reference to implications of lying.
0 – Reference to general non-specific information (because he looked everywhere else)

Strange Stories: Non Mental Story (White et al., 2009)

Sam decides to go on a long walk to get some fresh air. Unfortunately, just after leaving the house, the wind begins to pick up and it starts to rain. Luckily Sam always has an umbrella with him. He quickly puts up the umbrella and wraps his coat tightly around him. Suddenly a gust of wind blows the umbrella straight out of Sam's hand and it lands in a large, very prickly bush. Sam manages to run and fetch it before it blows off again and is pleased to find it all in one piece. As he walks home, he notices that his head is starting to get wet despite the umbrella.

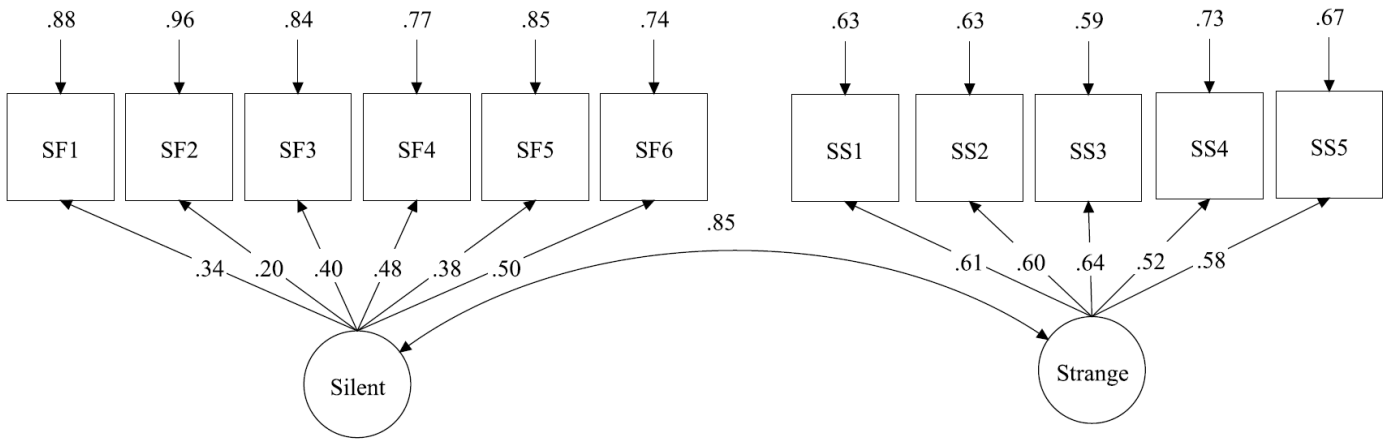
Why is Sam getting wet?
2 – Reference to the bush making holes in the umbrella
1 – Reference to either the bush or holes in the umbrella
0 – Reference to irrelevant or incorrect factors (it was raining, he hasn't got an umbrella)

Panel B. Sample Silent Film Task Item

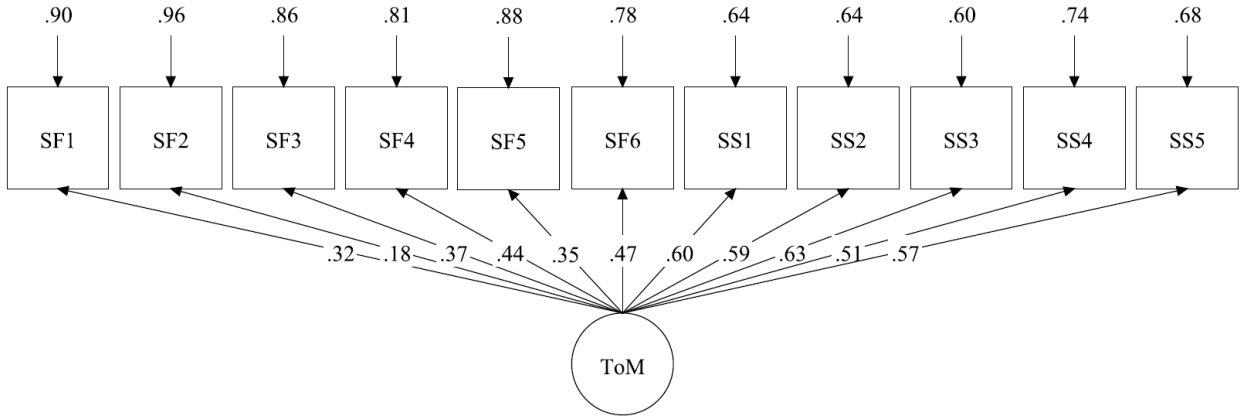
Screenshots taken with permission. From 'Safety Last' (Copyright of the Harold Lloyd Trust, 1923)
Why did the driver lock Harold in the van?
2 – The driver didn't know Harold was in the van; he didn't mean to.
1 – He wanted to continue on his rounds; He did not see/hear him.
0 – The man is deaf/hard of hearing (no reference to knowing); The man told him to; He kidnapped him.

Figure 2. *WLSMV Standardized Parameter Estimates for Theory of Mind Measurement Models*

Panel A. Two Latent Factor Model



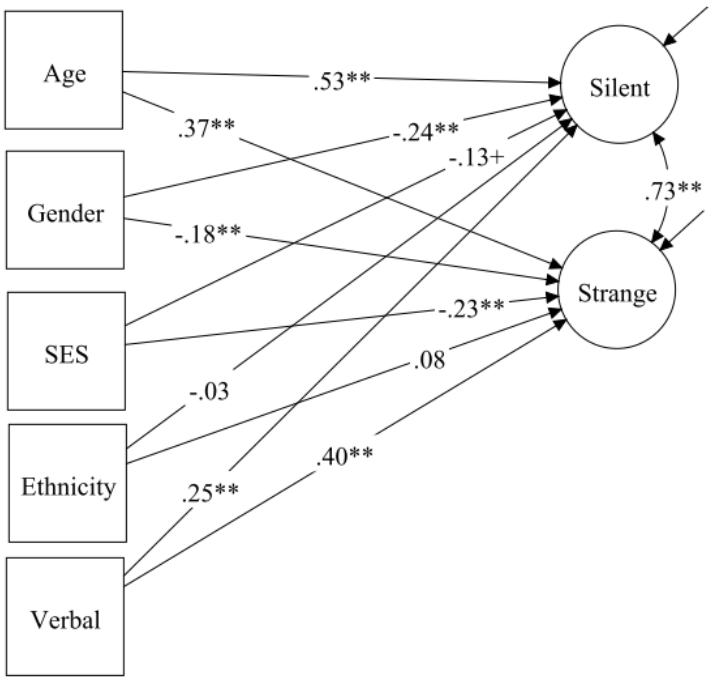
Panel B. Single Latent Factor Model



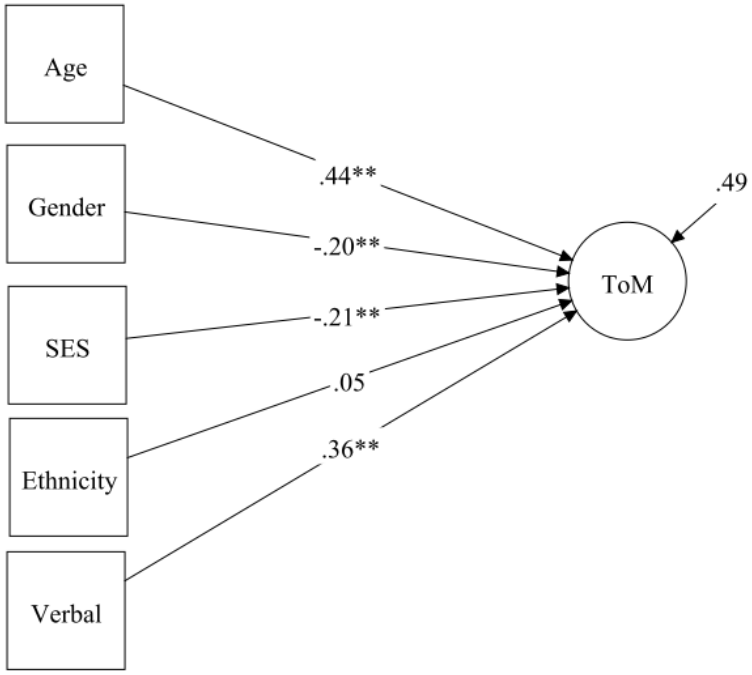
Note. ToM = Theory of Mind. SF = Silent Film Task. SS = Strange Stories Task (Mental State Items).

Figure 3. *WLSMV Standardized Parameter Estimates for Correlates of Theory-of-Mind Task Performance.*

Panel A. Correlates of Performance on the Strange Stories and Silent Film Tasks



Panel B. Correlates of Performance on the Theory-of-Mind Latent Factor



Note. $^{**}p < .01$. $^{*}p < .05$. $^{+}p < .10$.

Figure 4. *IRT Precision Estimates at Different Levels of Theory-of-Mind Latent Ability.*