Theory of Mind in Middle Childhood: Longitudinal Associations with Executive Function and Social Competence

Rory T. Devine, Naomi White, Rosie Ensor & Claire Hughes

University of Cambridge

Author Note

Center for Family Research, Department of Psychology, University of Cambridge, Free School Lane, Cambridge CB23RQ, United Kingdom. Rory T. Devine was funded by a Graduate Research Scholarship from St John's College, Cambridge. Naomi White was funded by a Rutherford Memorial International Scholarship from the Rutherford Foundation. Rosie Ensor was funded by a British Academy Post-Doctoral Fellowship. The first wave of data collection for this study was funded by a grant from the Economic and Social Research Council to Claire Hughes. We would like to acknowledge Keri Wong for her assistance with recruitment and data collection. Correspondence should be sent to Rory T. Devine via electronic mail: rtd24@cam.ac.uk

Key Words: Theory of Mind, Executive Function, Social Competence, Middle Childhood, Longitudinal.

Theory of Mind in Middle Childhood: Longitudinal Associations with Executive Function and Social Competence

Abstract

The vast majority of studies on theory of mind have focused on the preschool years. Extending the developmental scope of theory-of-mind (ToM) research presents opportunities to both reassess theoretical accounts of ToM and test its predictive utility. The twin aims of this longitudinal study were to examine developmental relations between ToM, executive function (EF) and teacher-rated social competence in middle childhood. 137 children (69 males) were followed across a 4-year period spanning middle childhood (*M* Ages at Waves 1 and 2 = 6.05, *SD*= .35, and 10.81, *SD* = .35). Individual differences in ToM were moderately stable across middle childhood. While there were concurrent associations between ToM and EF at both time points, there were no longitudinal links between these constructs. In contrast, there were concurrent and longitudinal links between ToM and teacher-rated social competence, such that individual differences in ToM predicted later social competence at school. These results are discussed in light of competing theories about the links between ToM and EF and the importance of individual differences in ToM for children's social lives.

Theory of Mind in Middle Childhood: Longitudinal Associations with Executive Function and Social Competence

Scientific interest in the human ability to reason about mental states and how children acquire a 'theory of mind' (ToM), has flourished over the past three decades, with lively debate still evident in developmental psychology and numerous related disciplines (for reviews see: Hughes & Devine, 2015; Wellman, 2014). Recent years have seen a massive expansion in the developmental scope of this research field, with exciting new findings from studies of infancy (e.g., Baillargeon, Scott & He, 2010), middle childhood (e.g. Devine & Hughes, 2013), adolescence (e.g., Dumontheil, Apperly & Blakemore, 2010) and adulthood (e.g., Duval, Piolino, Bejanin, Eustache & Desgranges, 2011). Nevertheless, an overwhelming majority of studies have continued to focus on the preschool years. A further limit to research in this area has been a heavy reliance on cross-sectional designs that preclude analysis of developmental relations. Addressing these twin gaps, the current longitudinal study examined developmental relations across middle childhood between individual differences in ToM and two key correlates: (i.) executive function (EF) and (ii.) social competence.

ToM and EF in Middle Childhood

Executive Function (EF) is an umbrella term that encompasses the set of higher-order cognitive abilities involved in controlling thought and action. These include over-riding habitual responses ('inhibition'), shifting between tasks ('cognitive set shifting') and holding information in mind to guide future action ('working memory') (Carlson, Zelazo & Faja, 2013). These processes are measured using tasks that appear quite distinct on the surface but are thought to measure a common underlying ability. The typically modest correlations between different measures of EF reflect the task impurity problem: differences in the non-

executive demands of tasks designed to measure EF (that is, the surface or task-specific features) are assumed to mask underlying commonalities (Miyake & Friedman, 2012; Miyake et al., 2000). To overcome this problem studies often adopt an aggregate approach to assessing EF via a number of different tasks designed to capture inhibition, set shifting, working memory or some combination thereof (Miyake & Friedman, 2012).

In a recent meta-analysis of data from 102 studies involving almost 10,000 3- to 6-year-old children from 15 countries, Devine and Hughes (2014) reported a moderate mean effect size for the correlation between one aspect of ToM (i.e., false belief understanding) and EF, $r = .38$. The goal of the meta-analysis was to test the predictions from competing theoretical accounts of the association between ToM and EF. According to the expression account, these constructs share overlapping peripheral task demands such that performance on measures designed to index ToM requires some degree of executive control (Perner, Lang & Kloo, 2002). In other words, EF is not developmentally linked with ToM. In contrast, others have proposed at least two distinct (and competing) *functional* relations between EF and ToM (Perner & Lang, 2000). For example, several theorists have posited that the acquisition of ToM is necessary for the development or emergence of EF (e.g., Carruthers, 1996; Perner & Lang, 1999). Specifically, understanding the relation between mental states and behavior (meta-representation) enables children to control their own thoughts and actions. Other theorists have argued that EF enables children to attend to and reflect upon others' mental states and so permits the growth of ToM (e.g., Russell, 1996). Supporting a hybrid 'emergence-expression' account, the meta-analysis revealed that the correlation between ToM and EF was stable in magnitude across age groups (from ages 3 to 6), different EF tasks and geographical regions but was weaker for indirect measures of ToM (i.e., tasks involving looking time responses) than for direct measures of ToM (i.e., tasks involving an

explicit verbal or motor response). The available longitudinal data showed an asymmetric developmental association between EF and ToM: while early EF predicted later ToM (even when controlling for previous ToM and verbal ability) early ToM did not predict later EF.

While much has been learned about the links between ToM and EF in the preschool years, relatively little is known about the association between these constructs in middle childhood. It is possible that the developmental association between ToM and EF might change beyond the preschool years; if only because middle childhood (the period from ages 6 to 11) is, in most cultures, characterized by entry into formal primary education. As a result children are exposed to increasingly sophisticated forms of knowledge (e.g., fictional literature), new expectations about regulating their behavior, and widening social horizons through interactions with peers and adults outside the family (Del Giudice, 2014). Alongside evidence for parallel age-related improvements across middle childhood and adolescence in both EF (e.g., Davidson, Amso, Anderson & Diamond, 2006; Huizinga, Dolan & Van Der Molen, 2006) and ToM (e.g., Apperly, Warren, Andrews, Grant &Todd, 2011; Devine & Hughes, 2013), recent cross-sectional studies show correlations between measures of these two constructs in middle childhood (e.g., Lagattuta, Sayfan & Blattman, 2010; Lagattuta, Sayfan & Harvey, 2014). However, the findings from these cross-sectional studies do not demonstrate that there are *developmental* links between EF and ToM beyond the preschool years.

The expansion of research on ToM into middle childhood (and beyond) raises important measurement questions, for which the competence/performance distinction (e.g., Meins, Fernyhough, Johnson & Lidstone, 2006) may be helpful. By middle childhood there is little doubt that children understand mental-state concepts such as beliefs and desires (i.e.,

ToM competence). Thus individual differences in middle childhood might reflect variation in children's ability to *use* this understanding flexibly and appropriately (i.e., ToM performance). Support for this proposal comes from recent studies (e.g., Apperly, Warren, Andrews, Grant &Todd, 2011; Banerjee et al., 2011) that show age-related differences in performance on the same task items across middle childhood on a range of ToM tasks that do not require any new conceptual insights. For example, Apperly et al. (2011) found age-related increases in performance across middle childhood on a task that required participants to judge which of two boxes a protagonist would look in for an object (i.e., to understand first-order false beliefs). As the authors suggest, it is unlikely that changes in children's conceptual understanding underpin such age-related contrasts.

The current longitudinal study applied age-appropriate batteries of tasks for each time-point (ages 6 and 10) that were designed to assess individual differences in children's ability to reason about others' beliefs, knowledge and desires. At the first time-point we used a task battery devised by Hughes et al. (2000), which has been shown to have good test-retest reliability and internal consistency and to be sensitive to individual differences in a sample of 1116 pairs of 5-year-old twins (Hughes et al., 2005). This battery consisted of picture-book and puppet-based tasks designed to measure first- and second-order false belief understanding as well as the ability to infer emotions linked to false beliefs. For the second time point, we constructed a novel multi-modal task battery that included two widely-used 'advanced' ToM tasks, the Strange Stories task (Happé, 1994) and the Triangles Task (Castelli et al., 2000), as well as one recently-designed task: the Silent Film task (Devine & Hughes, 2013). These three tasks used text-based vignettes, animated geometric figures and short clips from a silent film to depict social scenarios and children were required to refer to characters' mental states in order to explain their actions. Despite the growth in the number

and type of available 'advanced' ToM tasks suitable for use in middle childhood, little consideration has been given to the relations between tasks purporting to measure ToM in this developmental period. By including a battery of diverse ToM tasks we hoped to assess both these inter-relations and the links between performance on 'advanced' tasks and classic tests of first- and second-order false belief understanding.

Our first aim was to elucidate the nature of the links between individual differences in ToM and EF in middle childhood. To this end, we assessed children's performance in each domain at two time points spanning middle childhood at ages 6 and 10 (i.e., when children were in their first and final years of primary school). Our goal was to test the predictions from competing theoretical accounts based on findings from preschoolers. Specifically, the expression account predicts concurrent but not longitudinal links between these constructs. That is, while EF might be implicated in the *use* of ToM (and hence shows concurrent associations with ToM performance), it will not play a role in the development of ToM (Apperly, 2011). In contrast, two other accounts predict both concurrent and longitudinal associations between EF and ToM. For proponents of the meta-representation account (e.g., Perner & Lang, 1999), longitudinal associations would be expected between early ToM and later EF (only), while proponents of the emergence account (e.g., Hughes, 1998) would predict significant independent associations between early EF and later ToM (but not vice-versa). Meta-analytic evidence from preschoolers favors the latter account, but the nature of the association between ToM and EF has yet to be investigated in middle childhood.

ToM, EF and Social Competence in Middle Childhood

The vigor with which ToM has been researched over the past 30 years reflects a widely held assumption that ToM is a central foundation of everyday social interaction

(Ratcliffe, 2007). The growing recognition that typically developing children show striking variation in ToM led to speculation that these normative individual differences underpin key social outcomes in children's lives (Dunn et al., 1991). This drove the development of the 'social individual differences' account of ToM (Apperly, 2012). Although a large number of studies have demonstrated the influence of children's early social experiences on later individual differences in ToM (Hughes & Devine, 2015) there is less evidence to support the view that ToM matters for social competence. Findings from existing cross-sectional studies are mixed. Some studies show moderate to strong correlations between false belief understanding and adult ratings of young children's social competence (e.g., Watson, Nixon, Wilson & Capage, 1999) while others have failed to replicate these findings (e.g., Astington, 2003; Newton & Jenvey, 2011). In middle childhood, researchers have found weak to moderate links between ToM and peer-rated acceptance (Banerjee, Watling & Caputi, 2011) and self-rated loneliness (Devine & Hughes, 2013). These findings might indicate that individual differences in ToM matter more for children's social competence in middle childhood than in the early years.

A more critical examination of these findings exposes two shortcomings in the literature. First, Astington (2003) has highlighted the need to rule out potentially spurious associations between ToM and social competence reflecting, for example, shared links with language ability. Alongside language, links between EF and social competence have also been reported in early childhood (e.g. Razza & Blair, 2009; Rhoades, Greenberg & Domitrovich, 2009). Studies of the link between ToM and social competence in middle childhood have yet to account for the potentially confounding effects of EF. Second, the existing evidence in support of the social individual differences account is largely cross-sectional and so open to at least two possible explanations (Astington, 2003). That is, while

the ability to reason about others' mental states may impact on children's social behavior, it is equally possible that early acquisition of learned social behaviors facilitates children's performance on ToM tasks.

To date, longitudinal studies of the association between ToM and social competence in early and middle childhood have revealed mixed findings regarding the direction of association between these two constructs (e.g., Jenkins & Astington, 2000; Banerjee et al., 2011). Our second aim was therefore to examine the cross-sectional and longitudinal associations between ToM and teacher's ratings of children's social competence in school settings. We sought to extend the literature in two ways. Firstly, we controlled for potentially confounding factors such as individual differences in EF, verbal ability and socio-economic status in our analyses. Secondly, the availability of measures of ToM, EF and social competence at two time points permitted us to assess the direction of any developmental associations between these constructs across a 4-year period in middle childhood.

To summarize, our study had two primary aims. The first aim was to investigate whether EF and ToM continue to be developmentally (i.e., functionally) related across middle childhood, by examining cross sectional and longitudinal associations between EF and ToM. We were also able to gather teachers' ratings of social competence at each time-point and so our second aim was to examine the relations between individual differences ToM, EF and social competence in middle childhood.

Method

Participants

Participants were recruited from a panel of 184 children who had previously taken part in longitudinal study tracking children's social and cognitive development from ages 4 to 6 years. We achieved 90% contact (re-establishing contact with 166/184 families). Thirteen children were not eligible to take part, either because they no longer lived in the UK or because they had a diagnosed developmental disorder and a further 16 of those eligible either declined to take part or withdrew consent resulting in a final sample of 137 participants (69 males) (90% of eligible children). Binary logistic regression revealed that while non-returners did not differ significantly from returners in age, verbal ability or maternal education, they were more likely to be male, OR = 2.58, B = .95, SE = .42, p = .03 (See Table S1 for description of eligible and returning participants at Wave1). Although ethnically homogenous (94% white), the sample was socio-economically diverse: 36% of the participants' mothers had no educational qualifications or age 16 qualifications only. The mean age of participants at Wave 1 was 6.05 years, SD = .35, Range: 5.22 – 6.98. At Wave 2 the mean age of participants was 10.81 years, SD = .35, Range: 10.01 – 11.95. The mean testing interval was 4.76 years, SD = .38, Range: 3.27 – 5.57. Given the age spread in our sample, we controlled for age in our analyses.

Procedures

At both Wave 1 and 2, participants were tested individually at school by experienced graduate researchers using counter-balanced task batteries in which measures of EF alternated with measures of ToM (see Table 1 for summary of tasks). At the end of each session participants were thanked with a small gift or a book token. Teachers rated participants' social competence via a postal questionnaire. Teacher questionnaires were completed for 103 participants at both study waves. There were no differences between

participants whose teachers did or did not return questionnaires in terms of age or verbal ability at Wave 1, gender and maternal education, Nagelkerke $R^2 = .01$, χ^2 (4) = 1.08, $p = .89$. Importantly, there were no significant differences in teacher-rated social competence at Wave 1 between those whose teachers did ($M = 44.04$, $SD = 8.37$) and did not return ($M = 44.13$, $SD = 8.76$) questionnaires at Wave 2, t (127) = 0.05, $p = 0.96$.

Measures

ToM at Wave 1. In the *Object Transfer False-Belief Task* (Wimmer & Perner, 1983) participants watched while an object was moved from one location to another while a puppet was absent. The puppet then searched for the object in the original location. To succeed on this task participants needed to both explain the puppet's behavior and identify the object's actual location (control question). In the *Unexpected Contents False-Belief Task* (Bartsch & Wellman, 1989) participants were shown a cereal box and a plain box and asked where they would expect to find cereal. They were then shown that the cereal box was empty but the plain box contained cereal. To succeed on this task, participants needed to predict where an uninformed puppet would look for cereal and to recall where the cereal really was (control question). Scores from these two tasks were significantly correlated, ϕ (135) = .27, $p < .001$ and summed to create a 0-2 point 'Predict/Explain False Belief' score.

The *Emotion False-Belief Tasks* involved two puppet-based stories that included either a nice or nasty surprise (Harris et al., 1989). For example, in one story Leo the Lion received a nasty surprise when another character replaced the contents of a can of his favorite drink (Cola) with a drink he disliked (Juice). In both stories participants were only given test questions if they responded correctly to two forced-choice emotion comprehension questions (e.g., How does Leo feel when he drinks juice?). Participants were scored as failing these

items if they did not pass the two forced-choice comprehension questions. Test questions for both stories were also presented in a forced-choice format and required the participants to predict a character's false belief (e.g., What does Leo think is in the can, Cola or juice?) and recall the actual contents. Participants were also asked to infer an emotion based upon both belief and desire information (e.g., How does he feel before he drinks from the can?), and to explain their response (e.g., "He's happy because he thinks there's Cola in the can"). Finally, participants received two comprehension questions about the character's emotion on learning the actual contents of the container and were only credited with success if they passed both the relevant false belief prediction and comprehension questions. Thus across the two tasks participants' answered two questions in which they had to attribute a first order false belief. Scores on these two trials were correlated, ϕ (135) = .41, $p < .001$, and were summed together. We called this score 'Predict Other's False Belief (I)' (range: 0 – 2). Participants also answered two questions in which they had to infer and explain an emotion based on a false belief. The scores from these two trials were correlated, ϕ (135) = .47, $p < .001$, and summed to create the 'Infer/Explain Emotion based on False Belief' score (range: 0 – 4; that is, 1 point for each correct inference and explanation).

The *Second-Order False-Belief Tasks* assessed participants' ability to infer a character's beliefs about another character's beliefs via two picture book tasks (Sullivan et al., 1994). In one story, Peter's mother wants to surprise him for his birthday by buying him a puppy, but telling him she will buy a toy. Unbeknown to her, however, Peter discovers the puppy. Participants were first given a forced-choice test question about Peter's initial false belief and a control question about Peter's actual birthday gift. Following Peter's discovery of the puppy, participants were asked to predict (e.g., Peter's Granny asks Mum what Peter thinks he got for his birthday. What does Mum say?) and explain (e.g., "because Mummy

doesn't know that Peter saw the puppy") Peter's mother's second-order false belief and had to also pass two further control questions. Participants only succeeded on the first and second order false-belief test questions if they also passed the corresponding control questions. Scores on the first-order false belief question from each story were correlated, ϕ (135) = .51, $p < .001$, and so were summed together (range: 0 – 2). We called this score 'Predict Other's False Belief (II)'. Scores on the second-order false belief question from each story were correlated, ϕ (135) = .43, $p < .001$, and so were summed together. We called this final score 'Infer/Explain Second-Order False Belief' (range: 0 – 4). Sample scripts for these ToM tasks can be found in Hughes et al. (2000).

ToM at Wave 2. The *Strange Stories Task* (Happé, 1994) consisted of five audio-visual vignettes depicting social situations, each followed by an open-ended question that required participants to explain a character's behavior with reference to his/her mental states. One story involved a double bluff, two stories involved deception and two further stories involved misunderstandings. Happé's (1994) scoring scheme was used to code responses (For the full text of the Strange Stories task and corresponding marking scheme see: White, Hill, Happé & Frith, 2009). Participants received 0, 1 or 2 points for each incorrect, partially correct or fully correct response. A random selection of 25% of anonymized transcripts showed moderate to strong levels of inter-rater reliability, mean $\kappa = .70$, $.49 \leq \kappa \leq .87$, all *ps* $< .01$. Responses to each item were summed to create a 'Strange Stories' total score.

In the *Triangles Task* (For sample items and scoring guidelines see: Castelli, Happé, Frith & Frith, 2000) participants watched three short animations that featured two moving triangles and were designed to elicit mental state attributions (coaxing, teasing and surprising). After each clip, participants were asked to describe what had happened and their

responses were recorded, transcribed and rated for ascription of: (i) intentionality, with scores for each item ranging from 0 (non-deliberate actions) to 5 (deliberate actions with the goal of affecting others' mental states; and (ii) accuracy, scored as 0 (incorrect descriptions), 1 (imprecise or incomplete descriptions) or 2 (correct descriptions of the story or actions presented in the clip). In a random selection of 25% of the anonymised transcripts, intentionality ratings for individual items exhibited strong inter-rater reliability, mean ICC = .80, range .69 - .90, all *ps* < .01, as did accuracy ratings, mean $\kappa = .71$, $.60 \leq \kappa \leq .80$, all *ps* < .001. We summed together intentionality and accuracy scores for each item to create a 'Triangles' total score.

In the *Silent Film Task* (For test description and sample items see: Devine & Hughes, 2013) participants were required to explain the behavior of characters in scenarios presented in five brief clips from a classic silent comedy. The clips were played once in a fixed order. The first clip was accompanied by two questions and the remaining four clips were followed by one question each. The responses were scored using the scheme developed by Devine and Hughes (2013): participants received 2 points for responses that provided an accurate mentalistic explanation of events (e.g., The driver didn't *know* that Harold was in the van), 1 point for responses that refer to the facts but don't use mentalistic explanations (e.g., The driver did not hear Harold) and 0 points for irrelevant or incorrect responses (e.g., The driver kidnapped Harold). In a random selection of 25% of the anonymized transcripts individual items showed moderate to strong agreement, mean $\kappa = .74$, $.47 \leq \kappa \leq .93$, all *ps* < .01. Scores across each of the 6 questions were summed together to create a 'Silent Film' total score. [1]

[1] Participants also completed the automated Director Task (Dumontheil et al., 2010). This task did not load onto the ToM latent factor but instead loaded onto the EF latent factor. In addition this task did not correlate significantly with Wave 1 measures of ToM. The results of this task have been omitted but are available from the first author.

EF at Wave 1. Participants completed the *Tower of London Task* (Shallice, 1982) using simplified age-appropriate instructions (Hughes, 1998). This task challenges both inhibition (e.g., over-riding the impulse to start moving the balls) and working memory (e.g., holding in mind the three rules and the target arrangement). Participants had to move a set of three colored balls placed in a starting arrangement on three wooden dowels to each target arrangement within a specified number of moves. They were told that (1) only one ball could be moved at a time, (2) balls underneath another ball could not be moved and (3) the maximum number of balls that each dowel could hold (i.e., 3, 2, 1). Trials were administered in a fixed order and consisted of three 2-move problems, three 3-move problems and three 4-move problems. Participants received 1 point for each perfect solution and 0 points for solutions that exceeded the number of permissible moves (possible range: 0 – 9).

Participants completed the *Bead Memory task* from the Stanford-Binet Intelligence Scale (Thorndike, Hagan & Sattler, 1986) at Wave 1 and Wave 2. On each trial participants were shown for 5 seconds an arrangement of beads on a stick that varied in shape (ellipsoids, spheres, cones and disks) and color (red, white and blue). They were then given a box of beads (with three of each type) and asked to reproduce each arrangement exactly. The task was discontinued after 3 failures across 4 trials; raw scores were calculated by subtracting the number of failed trials from the highest item attempted. Although initially designed to measure short-term visual memory, this task challenges multiple EFs (e.g., participants must not touch the beads while the image is displayed, the image is then removed such that they must hold it in mind while they select the beads, they must then plan ahead in order to place the beads on the stick in the correct sequence). Performance on this task is correlated with performance on measures of inhibition and set shifting in childhood (e.g., Hongwanishkul, Happaney, Lee & Zelazo, 2005; Hughes & Ensor, 2005).

The *Day/Night Test* of conflict inhibition (Gerstadt, Hong & Diamond, 1994) had 12 trials; on each trial participants were shown a picture of the sun or moon and asked to say 'night' in response to the sun and 'day' in response to the moon. Given the clear significant negative skew in the distribution for this task, $M = 11.12$, $SD = 1.85$, $Z_{skew} = -15.38$, $p < .001$, we dichotomized scores on this task using a median split: participants scoring fewer than 12 trials out of 12 correctly ($N = 45$) received a score of 0 and those scoring 12 correctly received a score of 1. The mean accuracy score for the 'failing' group was 9.33, $SD = 2.39$.

The *Trucks Game* tested participants' cognitive set shifting (Hughes & Ensor, 2005). On the first trial of the initial rule-learning phase of this game, the child was shown a picture of two trucks on yellow card and asked to guess which truck would win him or her a treat (in fact, the child's choice was always judged as the correct choice and the child was told to remember that truck). On the second trial, the child was shown a second pair of trucks and asked to select one of the trucks; again this choice was always recorded as correct and the child was told to remember that truck too. On the next 6 trials the child was shown either the first or second pair of trucks in a pseudo-random order with the spatial position (left or right) of each truck counter-balanced. Within each pair the trucks were quite similar, such that some children chose the wrong truck by mistake on early trials; they were therefore given verbal feedback on each trial and categorized as passing the pre-shift phase if they chose the correct truck for at least four out of the final five trials. Next children completed the post-shift phase. This time, each pair of trucks was presented on a green card and the child was told to select the other (previously non-rewarded) truck in order to win a treat. There were 8 post-shift trials and success was indexed by correct choices on at least 4 of the last 5 trials. We awarded a score of 1 to participants passing both phases ($N = 121$) and 0 ($N = 16$) to those who did not.

EF at Wave 2. To measure conflict inhibition, participants completed the *Arrows Task* (Davidson et al. 2006). In this task participants viewed one of four images of a purple arrow on a white screen. The arrow pointed either directly downward or diagonally on either the left- or right-hand side of the screen. During congruent (control) trials, the arrow pointed directly downward and participants had to press the button on the same side as where the arrow appeared. During incongruent (test) trials, the arrow pointed diagonally and participants had to press the button on the opposite side to where the arrow appeared. Before beginning, participants received detailed instructions and practice items. Each 750*ms* trial was preceded by a 500*ms* interval in which a 16-point font black crosshair was displayed against a white background. Participants had to respond with a button press within the 750*ms* trial period to each of 12 control trials and 12 test trials administered in a random order. Efficiency scores were based on the total number of correct incongruent trials divided by the total time taken on incongruent trials. In line with previous studies (Davidson et al., 2006) anticipatory responses of < 200*ms* and missed responses were recorded as errors.

The *Smiling Faces Task* (Huizinga et al., 2006) is an index of set shifting. During this task participants had to respond to one of four cartoon faces (i.e., a happy boy, a happy girl, a sad boy and a sad girl) displayed in one of four quadrants on a screen. In single task trials participants were asked to indicate whether a face was (1) a boy or a girl if it appeared in the top two quadrants or (2) happy or sad if it appeared in the bottom two quadrants. During alternating trials, participants had to integrate both rules. Trials were administered in four blocks in a fixed order: one set of 16 single task trials (boy or a girl); another one set of 16 single task trials (happy or sad); and two sets of 16 alternating trials. Each trial lasted 3500*ms* during which time participants had to respond with a button press. The trials were preceded by a 500*ms* interval in which a 16-point font black crosshair was displayed in the center of

the screen on a white background. Trials were presented in a random order within each block. We calculated efficiency scores by dividing the total number of correct alternating trials by the total time taken on alternating trials.

Verbal Ability. We used the *British Picture Vocabulary Scale* (BPVS) (Dunn, Dunn, Whetton & Burley, 1997) to measure verbal ability at both waves. On each item of this receptive vocabulary task (suitable for children aged 3- to 16 years) participants were asked to pick one picture from a set of four that matched a word read aloud by the examiner. We subtracted the number of errors from the item number corresponding to the last number in the participant's ceiling set (that is, sets containing 8 or more errors) to calculate raw scores.

Social Competence at School. At both study waves teachers completed the 30-item *Social Skills Rating System* (SRSS) (Gresham & Elliot, 1990). The items contain examples of behavior associated with cooperation, peer relations, adherence to social conventions, assertion, and self-control in social situations at school. Teachers rated the frequency of each behavior on a 0 /1 /2-point scale (i.e., Never, Sometimes, Very Often). Total scores therefore ranged from 0 to 60 (Wave 1 $\alpha = .90$; Wave 2 $\alpha = .92$).

Results

Analytic Strategy

Latent variable and structural equation modelling were undertaken using *MPlus* Version 7 (Muthèn & Muthèn, 2012). To assess the coherence of our diverse tests of ToM and EF we conducted a confirmatory factor analysis (CFA), using a mean- and variance-adjusted weighted least squares estimator to generate model parameters when task scores were categorical (see Table 1) and full information maximum likelihood estimation when the

data were continuous and normally distributed (so that all cases with data in Wave 1 could be included in both sets of analyses). To examine longitudinal associations in our data, we used structural equation modelling. We used four primary criteria to assess the acceptability of our measurement and structural models: a non-significant χ^2 test of model fit, Comparative Fit Index (CFI) > .95, Tucker Lewis Index (TLI) > .95 and Root Mean Square Error of Approximation (RMSEA) < .08 (Brown, 2006).

To reduce the number of parameters in our structural equation models while capitalizing on the benefits of using latent factors, we used the Wave 1 and Wave 2 factor scores in the longitudinal models (see Data Reduction). In each of the longitudinal models we allowed the predictor variables to inter-correlate. Consistent with previous studies on the longitudinal links between EF and ToM (e.g., Hughes & Ensor, 2007), we sought to examine the independent association between the Wave 1 independent variable and the Wave 2 dependent variable controlling for Wave 1 scores on the dependent variable, maternal education and Wave 1 verbal ability. We adopted a similar approach when examining the longitudinal links between ToM, EF and teacher-rated social competence.

For the EF and ToM variables at Wave 1 and Wave 2, missing data did not exceed 1% on any variable in the dataset. No participant had missing data for all test variables. Note that there were no 'gateway' items in our data. That is, if a participant failed a control question, the participant received a score of '0' for that item. There were no missing values on any of the background variables (i.e., gender, age, verbal ability, maternal education). For teacher-rated social competence, 129 teachers provided ratings at Wave 1 and 103 teachers provided ratings at Wave 2. To avoid loss of data, missing values were estimated in *Mplus* using either maximum likelihood or mean- and variance- weighted least squares depending

on the model estimator (Muthèn & Muthèn, 2010). *Mplus* provides reliable estimates of missing values and latent variable factor scores based on participants' scores on other variables in the model (Asparouhov & Muthèn, 2010a).

Data Reduction

Tables 2 and 3 show the descriptive statistics (and correlations) for ToM and EF scores respectively. The mean score on the teacher-rated SRSS was 44.04, *SD* = 8.37, Range: 18 – 60 at Wave 1 and 42.72, *SD* = 9.18, Range: 20 – 60 at Wave 2. In line with previous studies of individual differences in EF using confirmatory factor analysis, zero-order correlations between the EF tasks at both time points were weak to moderate in strength (e.g., Miyake et al., 2000). Likewise the inter-correlations between the ToM scores at both time points were also weak to moderate in strength. CFA can be used in such situations to extract a 'pure' measure of each construct by statistically isolating the shared variance between the diverse measures of each construct from the task-specific variance and measurement error (Brown, 2006; Miyake et al., 2000).

For the Wave 1 data, using CFA, we specified two measurement models that replicated the latent factors for ToM and EF reported elsewhere in the literature (Hughes, Ensor & Marks, 2011; Hughes & Ensor, 2011). Standardized factor loadings and residuals for each of these models are depicted in Figure 1 panels A and B. First we specified a model in which each of the false-belief scores loaded onto a single ToM latent factor. We set the metric of the latent variable by fixing the loading of the marker indicator (i.e., Predict/Explain False Belief) to 1 (Kline, 2011). This model provided an excellent fit to the data, χ^2 (4) = 2.71, p = .61, CFI = 1.00, TLI = 1.00, RMSEA = 0. The latent factor exhibited significant variance, *Unstandardized Est*. = 0.21, p = .04. Next we specified a single factor

model in which each of the EF task scores loaded onto a single latent factor. Once again we set the metric of the latent variable by fixing the loading of the marker indicator (i.e., Tower of London) to 1. The model provided an excellent fit to the Wave 1 EF data, χ^2 (2) = 1.91, p = .38, CFI = 1.00, TLI = 1.01, RMSEA = 0. The latent factor exhibited significant variance, *Unstandardized Est*. = .09, p = .01.

Next we specified two latent factor models to measure individual differences in ToM and EF at Wave 2. The standardized factor loadings and residuals for each of these models are depicted in Figure 1 panels C and D. These models were 'just-identified' since there was an equal number of model parameters and variances/covariances in the sample matrix (Brown, 2006). While model fit estimates cannot be calculated for just-identified models because there are 0 degrees of freedom, parameter estimates can be still calculated and interpreted (Brown, 2006). In the first model each of the ToM scores loaded onto a single ToM latent factor. We set the metric of this latent factor by fixing the loading of the Strange Stories task to 1. This latent factor exhibited significant variance, *Unstandardized Est.* = 2.82, $p < .05$. In the second model each of the EF measures loaded onto a single EF latent factor. We set the metric of this latent factor by fixing the Bead Memory task to 1. This latent factor showed significant variance, *Unstandardized Est.* = 3.88, $p < .05$.

We created factor scores for each of the latent variables for ToM and EF by imputing plausible values (i.e., a distribution of factor scores for each participant) for each latent variable using Bayesian estimation in *Mplus* (Asparouhov & Muthèn, 2010b). Standard factor scores calculated using the regression method do not provide reliable estimates of true scores when used as dependent variables in regression models (Skrondal & Laake, 2001). Plausible values, however, can be used to build models for secondary analysis and provide reliable

estimates of the true scores on latent factors (Asparouhov & Muthèn, 2010b). We used plausible values based on multiple imputations in each of the longitudinal structural equation models reported below.

To examine the convergent validity of Wave 2 ToM measures we assessed the correlation between the scores from the ToM latent factor from Wave 1 and performance on each of the measures of ToM at Wave 2. Performance on the ToM latent factor at Wave 1 was correlated with performance on the Strange Stories, r (135) = .41, p < .001, 90% CI [.29, .52], the Triangle Task, r (135) = .27, p < .01, 90% CI [.14, .40] and the Silent Films Task, r (135) = .35, p < .001, 90% CI [.23, .45]. When concurrent age and Wave 2 verbal ability were accounted for, the Strange Stories, pr (133) = .27, p < .01, 90% CI [.16, .39], Triangles Task, pr (133) = .22, p < .05, 90% CI [.08, .36], and Silent Film Task, pr (133) = .25, p < .01, 90% CI [.12, .40] remained correlated with Wave 1 ToM.

ToM and EF in Middle Childhood

Our first aim was to examine the associations between EF and ToM within and across each wave of the study. As shown in Table 4, there were moderate and significant correlations within and across waves. Although attenuated, at both time-points concurrent associations between ToM and EF remained significant even when concurrent verbal ability and age were taken into account. To assess the developmental relations between these two constructs across middle childhood we specified a longitudinal structural equation model (see Figure 2 for path diagram with standardized estimates and Table S2 for unstandardized parameter estimates). We regressed ToM scores at Wave 2 onto ToM scores at Wave 1, EF scores at Wave 1, verbal ability at Wave 1 and maternal education (i.e., No qualifications or GCSE only = 0; A-Level or Degree = 1). We also regressed EF scores at Wave 2 onto EF

scores at Wave 1, ToM scores at Wave 1, verbal ability at Wave 1 and maternal education. We permitted each of the variables at Wave 1 and Wave 2 to inter-correlate. This cross-lagged model fit the data well, χ^2 (3) = 2.84, p = .42, CFI = 1.00, TLI = 1.00, RMSEA = 0. As shown in Figure 2 (and Table S2), there was weak but significant stability in individual differences in EF and ToM across middle childhood. However, there were no significant cross-lagged associations between EF and ToM across middle childhood.

We performed a post-hoc power analysis using Monte Carlo simulation in *Mplus* 7. Using the means and variances from each variable in our sample, we tested whether a sample size of 137 participants would be sufficient to detect small-to-medium effects in the cross-lagged paths of the model described above. We specified the same cross-lagged model but calculated the unstandardized regression co-effecients for the two cross-lagged paths in our model if the correlation was .30. According to Muthèn and Muthèn (2002) sample sizes should be large enough to provide unbiased parameter estimates, unbiased standard errors (so that the significance of effects are not over- or under-estimated), sufficient coverage (that is, the proportion of replications for which the 95%CI contains the population value should be ≥ .91) and sufficient power (that is, the proportion of replications for which the null hypothesis of no relationship is rejected should be ≥ .80). Focusing on the two cross-lagged model parameters, we assessed four key features, recommended by Muthèn and Muthèn (2002), to determine whether the sample size of 137 was sufficient: parameter bias < 10%; standard error bias < 5%; coverage between .91 and .98; power close to .80. Our results revealed that both paths exhibited low levels of parameter bias (both < 1%) and low levels of standard error bias (both <5%). Coverage for both parameters was .94 and the power to detect medium effects for both parameters was close to .80 (.78 and .84 respectively). Together these results suggest that a sample of 137 participants was sufficient to detect medium effects.

ToM, EF and Social Competence in Middle Childhood

Our second aim was to assess the links between ToM, EF and teacher-rated social competence at school. Note that some teachers provided questionnaires for more than one study participant. Specifically, 13 teachers completed questionnaires on 2 participants, 3 teachers on 3 participants, 1 teacher on 4 participants, 1 teacher on 5 participants and 1 teacher on 6 participants. The remaining questionnaires were completed independently. Given that the data were clustered and so variance in social competence ratings could be attributable to both child-level variation and teacher-level variation, we calculated an intra-class correlation (ICC) to determine the proportion of total variance in social competence ratings that was due to between-teacher variance (Byrne, 2012; Muthèn, 1997). The ICC for the social competence rating was .11 indicating that between-teacher variance was minimal (Byrne, 2012). We therefore assessed the longitudinal relations between ToM, EF and social competence across middle childhood using a single-level longitudinal structural equation model (See Figure 3 for path diagram with standardized estimates and Table S3 for unstandardized parameter estimates). Specifically, Wave 2 ToM scores were regressed onto Wave 1 ToM scores, Wave 1 EF scores, Wave 1 teacher-rated social competence scores, Wave 1 verbal ability, gender (i.e., Male = 0; Female = 1) and maternal education. Wave 2 teacher-rated social competence scores were regressed onto Wave 1 teacher-rated social competence scores, Wave 1 ToM scores, Wave 1 EF scores, Wave 1 verbal ability, maternal education, and gender. Finally Wave 2 EF scores were regressed onto Wave 1 EF scores, Wave 1 ToM scores, Wave 1 teacher-rated social competence scores, Wave 1 verbal ability, gender and maternal education. Each of the variables at Wave 1 and Wave 2 were permitted to inter-correlate with other concurrent measures. This cross-lagged model provided an excellent fit to the data, χ^2 (5) = 3.37, p = .64, CFI = 1.00, TLI = 1.00, RMSEA = 0.01.

Inspection of this model revealed that individual differences in ToM, EF and teacher-rated social competence exhibited modest but significant relative stability across middle childhood (see Figure 3). Importantly, Wave 1 ToM scores showed a weak but significant independent association with Wave2 teacher-rated social competence at school. In contrast, Wave 1 teacher-rated social competence was unrelated to Wave 2 ToM scores. Moreover there were no cross-lagged associations between EF scores and teacher-rated social competence scores. To confirm our findings we re-ran this model using a multi-level framework clustering the social competence data by teacher. The model provided an excellent fit to the data, χ^2 (6) = 3.37, p = .76, CFI = 1.00, TLI = 1.00, RMSEA = 0.01 and the pattern of results remained unchanged.

We performed a Monte Carlo simulation to examine whether the sample size of 137 provided sufficient power to detect medium effects (r = .30) on each of the six cross-lagged paths in the model. As before we used the means and variances for each variable from our sample and specified a model identical to that described above. In addition we specified that data on the teacher-rated social competence measure at both time points would be missing for approximately 25% of the sample. For each of the 6 cross-lagged paths there were low levels of parameter bias (0.11 – 1.79%) and acceptable levels of standard error bias (2.07% - 4.88%). The coverage ranged from .93 to .94 and power estimates ranged from .77 to .85. These findings suggest that the sample size provided sufficient power to detect small to medium effects in the specified model.

Discussion

ToM and EF in Middle Childhood

Our study revealed that, despite the differences in the ToM tasks used at ages 6 and 10, both time-points showed similarly moderate *concurrent* associations between individual differences in ToM and EF that echoed the findings from a meta-analysis of preschool data (Devine & Hughes, 2014). In contrast, there were no significant cross-lagged associations between these two constructs across the 4-year interval between study time-points. This lack of longitudinal association is open to at least three possible explanations.

First, aspects of the study design might have masked any longitudinal effects. Specifically, the extended measurement interval (which exceeded that of all previous longitudinal studies of ToM and EF) might have attenuated the association between individual differences in ToM and EF. In addition some of our measures of EF at Wave 1 may not have been sufficiently challenging for the participants as indicated by evidence of ceiling effects on the Day/Night Task (67% at ceiling) and the Trucks Game (88% at ceiling). The other two EF tasks at Wave 1 were developmentally appropriate and exhibited normal distributions. That said it is still possible that ceiling effects in two of our EF measures might have attenuated any associations between EF and ToM. Finally, although the post-hoc power analyses revealed that there was a sufficiently large sample size to detect medium cross-lagged effects, it is still possible that our sample was not large enough to detect small cross-lagged effects. Larger samples would be needed to detect such effects.

The second possible explanation is that the presence of concurrent but not longitudinal associations between EF and ToM might indicate that these two constructs develop in tandem across middle childhood. Consequently, the observed asymmetry in the longitudinal links between EF and ToM in early childhood (Devine & Hughes, 2014) might disappear while concurrent associations might remain. The third possibility is that EF is

related to ToM performance in middle childhood but may not drive any further *developmental* change in ToM in this period. If so, this would suggest a developmental shift in the nature of the relation between these two constructs from emergence (in the preschool years) to expression (in middle childhood).

Note, however, that the lack of developmental links between ToM and EF in middle childhood does not mean that the concurrent associations between these constructs are trivial. Although gains in EF across middle childhood may have no implication for the continued growth of ToM across this period, EF might still play an important role in the process of reading others' minds. Concurrent correlations between ToM and EF in middle childhood and beyond indicate that EF is an integral component of ToM beyond the preschool years. As Apperly (2011) has proposed, EF may contribute to many of the processes involved in attributing mental states to others, including inferring what others think, temporarily storing this information in mind and using this information to explain and predict others' behavior.

Given that there are at least three potential explanations for our findings, our study signals the need for further longitudinal research on the links between EF and ToM in middle childhood and more specifically the need for repeated-measures designs over shorter measurement intervals. The extended temporal span of the current study precluded us from using the same measures of ToM and EF at both time points in order to examine directly whether *change* in one construct results in change in the other. In recent years a wide range of tasks suitable for measuring ToM (and indeed EF) across middle childhood has become available. Designs incorporating multiple time points in middle childhood will reveal whether changes in EF are related to growth in ToM and vice versa. Short-term repeated-measures studies could be used to test whether EF is implicated in more proximal developmental

changes in ToM in middle childhood. Finally, looking beyond longitudinal associations between EF and ToM, the use of tasks designed to isolate particular components of reasoning about others' mental states (e.g., Apperly, Back, Samson & France, 2008) may lead to greater insights into the role of EF in the expression or use of ToM beyond the preschool years.

ToM, EF and Social Competence in Middle Childhood

Our moderate within-time correlations between ToM and teachers' ratings of social competence at ages 6 and 10 complement previously reported associations between individual differences in ToM and self- and peer-reported social competence in middle childhood (e.g., Banerjee et al., 2011; Devine & Hughes, 2013). Teachers are well placed to provide more objective information about children's social behavior relative to others of a similar age (e.g., Stone, Otten, Engels, Vermulst & Janssens, 2010) and the convergence of findings from different sources supports the relevance of ToM to children's social lives in middle childhood. Taken together with existing peer-rated and self-reported data, our findings suggest that individual differences in ToM in middle childhood might be important for successful social interaction and behavior at school. The unidirectional nature of the association between early ToM and later social competence in middle childhood supports the claim that individual differences in ToM are socially meaningful (e.g., Astington, 2003) but contrasts with reports of reciprocal relations between ToM and social competence across a 1-year measurement interval (e.g., Banerjee et al. 2011). An extended interval of time might be required to detect the direction of the relation between ToM and social competence. It is also possible that peer acceptance might provide quite a distinct measure of social competence than teachers' ratings of social behavior in the classroom.

In contrast, we found only weak concurrent correlations and no evidence of cross-lagged associations between EF and social competence between the ages of 6 and 10. Studies have reported mixed findings about the links between performance on measures of EF and teacher- or parent-rated social competence in early and middle childhood. While there appear to be weak concurrent associations between EF and social competence in early childhood (e.g., Rhoades, Greenberg & Domitrovich, 2009), these findings have not been supported by longitudinal studies (e.g. Hughes & Ensor, 2011; Razza & Blair, 2009). Moreover, in a recent study by Harms, Zayas, Meltzoff and Carlson (2014), individual differences in EF at age 8 and 12 were unrelated to teachers' ratings of children's social competence at age 12. That said, there is a growing body of evidence to suggest that low levels of EF are related to problem behaviors (e.g., Espy, Sheffield, Wiebe, Clark & Moehr, 2011) and to academic underachievement (e.g. McClelland, Cameron, Connor, Farris, Jewkes & Morrison, 2007). It appears that while EF and ToM are correlated in early and middle childhood, individual differences in these abilities contribute to different developmental outcomes. Further longitudinal research will be needed to test this hypothesis.

Our findings extend the current literature both by documenting the developmental reach of early individual differences in ToM and by demonstrating the independence of longitudinal links between ToM and later social competence at school from other factors (e.g., verbal ability, maternal education and EF). This second point suggests that the determinants of social competence may differ in early and middle childhood. In contrast with findings from the preschool years (e.g., Astington, 2003), verbal ability was unrelated to teacher's ratings of social competence at age 10. Given the inconsistent findings about the links between ToM and different aspects of social competence in the early years, our findings suggest that individual differences in ToM become more salient in middle childhood. That

said, our reliance on a single measure of verbal ability (and indeed social competence) precludes firm conclusions about the relations between language and social competence.

Likewise, our findings do not provide unqualified support for the notion that ToM is vital for children's social lives. The data indicate that while ToM clearly plays a part in children's social competence at school in middle childhood, it is by no means a sufficient explanation (Astington, 2003). Consistent with previous findings about the links between ToM and social competence in middle childhood, the longitudinal association reported here, although independent and significant, was small in magnitude. Together, these findings highlight the importance of other unmeasured influences on children's social competence. Moreover the data reported here are correlational in nature. Training studies would provide more robust evidence for claims regarding the developmental relations between ToM and social competence both in the preschool years and in middle childhood. In one recent study of 9- and 10-year-old children, Lecce, Bianco, Devine, Hughes and Banerjee (2014) found that children in a 2-week conversation-based ToM training program showed greater gains in performance on the Strange Stories task than children assigned to an active control group (in which children discussed physical events in short vignettes). Further work is needed to establish whether the benefits of training for ToM performance result in gains in children's social competence in school and other settings.

Implications for Understanding Individual Differences in ToM

This study showed, for the first time, that performance on classic false-belief tasks predicted later performance on a battery of 'advanced' ToM tasks, such that individual differences in ToM exhibited weak to moderate stability from ages 6 to 10. Individual differences in ToM have been conceptualized in at least two ways. Some researchers have

interpreted individual differences in children's ToM task performance as either developmental delays or precocities (e.g., Slaughter & Repacholi, 2003; Wellman, Cross & Watson, 2001). Underlying this 'developmental lag' perspective is the assumption that children who perform poorly on false-belief task batteries will eventually 'catch up' with their peers with no real developmental consequences (Bartsch & Estes, 1996). In contrast, others have argued that variation in the age at which children obtain an understanding of mental-state concepts might predict later variation in ToM performance, that is, the ease and/or accuracy with which individuals attribute mental states to others (Apperly, 2012). Both the stability of individual differences in ToM across the 4-year interval in this study and the predictive association between early ToM and later teacher-rated social competence support this 'genuine variation' model.

Second, our longitudinal findings suggest that variation in cognitive skills such as verbal ability and EF do not fully explain individual differences in ToM performance in middle childhood. The stability of individual differences in cognitive abilities might reflect stability in either genetic or environmental influence (Tucker-Drob & Briley, 2014). Previous cross-sectional twin studies of ToM in 5-year-olds (Hughes et al., 2005) and 9-year-olds (Ronald et al., 2006) demonstrate very little genetic influence on individual differences in ToM but do not rule out the possibility of genetic influence on the stability of individual differences in ToM across middle childhood and beyond. Our study findings are also silent on this issue as the two time-points both fell within a developmental period when British children attend primary school and so typically remain with the same group of classmates. Recent longitudinal work in middle childhood suggests that peer acceptance predicts later individual differences in ToM (Banerjee et al., 2011). Therefore, if the environmental

account of stability is correct, one might predict a decrease in the longitudinal stability of ToM across the transition to secondary school when children enter a new social milieu.

Earlier we noted that recent studies have demonstrated age-related improvements in performance on a range of 'advanced' ToM tasks across middle childhood. Researchers must now attempt to identify what factors might underpin the continued growth of the ability to reason about others' mental states in this developmental period. Social factors may have a main effect on the development of ToM beyond the preschool years. To date research on ToM, both in the preschool years and beyond, has focused on social or cognitive correlates in isolation. If cognitive factors, such as EF, do not exert a main effect on the development of ToM across middle childhood, these factors could nonetheless act as moderators of social influences. Enhanced EF, for example, might enable children to better attend to and reflect upon relevant social experiences (Hughes & Devine, 2015).

For more than three decades, research on ToM has focused predominantly on the preschool years. In this paper we have attempted to extend the traditional developmental scope of ToM research by examining the links between ToM, EF and social competence in middle childhood. Our findings revealed concurrent but not longitudinal associations between individual differences in ToM and EF across a four-year period in middle childhood. In addition, our findings showed both concurrent and longitudinal correlations between individual differences in ToM and children's social competence at school. These findings both provide an opportunity to re-evaluate existing theories of ToM and suggest promising future avenues for future research on individual differences in ToM in middle childhood.

Table 1. *Summary of Key Measures at Wave 1 and Wave 2*

Construct	**Task**	**Wave**	**Ability Assessed**	**Scale**
ToM	Object Transfer & Unexpected Contents	1	Predict and explain a character's behavior with reference to false beliefs.	Categorical
	Emotion False-Belief Task Puppet Stories	1	Attribute a false belief to a character. Infer and explain a character's emotion based on his/her false beliefs.	Categorical
	Second-Order False-Belief Task	1	Attribute a false belief to a character. Infer and explain a character's emotion based on his/her false beliefs.	Categorical
	Strange Stories Task	2	Explain a character's behavior with reference to mental states across five short text-based vignettes.	Continuous
	Triangles Task	2	Attribute mental states to explain the actions of two animated triangles.	Continuous
	Silent Film Task	2	Explain a character's behavior with reference to mental states across five clips from a silent film.	Continuous
EF	Tower of London	1	Planning, working memory, inhibition	Continuous
	Bead Memory Task	1 & 2	Working memory, inhibition	Continuous
	Day/Night Task	1	Inhibition	Categorical
	Trucks Game	1	Set shifting	Categorical
	Arrows Task	2	Inhibition	Continuous
	Faces Task	2	Set shifting	Continuous
Verbal	BPVS	1 & 2	Receptive vocabulary	Continuous

Table 2. *Descriptive Statistics and Correlations for Theory of Mind Tasks*

	Task	Wave 1					Wave 2		
		1	2	3	4	5	6	7	8
1	Predict/Explain False Belief	-	.26**	.20*	.24**	.24**	.18*	.10	.14
2	Predict False Belief (I)	.24**	-	.25**	.51**	.25**	.27**	.16	.17*
3	Predict False Belief (II)	.14	.17*	-	.37**	.79**	.35**	.06	.25**
4	Infer/Explain Emotion based on False Belief	.17*	.46**	.26**	-	.40**	.33**	.32**	.24**
5	Infer/Explain Second-Order False Belief	.18*	.14	.75**	.25**	-	.38**	.09	.30**
6	Strange Stories	.10	.20*	.24**	.20*	.26**	-	.34**	.31**
7	Triangles Task	.07	.12	.03	.29**	.03	.24**	-	.21*
8	Silent Film Task	.12	.10	.17*	.12	.20*	.22**	.17*	-
	Mean	1.50	1.84	1.09	2.90	1.50	6.39	12	7.66
	SD	0.66	0.46	0.86	1.53	1.59	2.39	3.97	2.11
	Range	0 – 2	0 – 2	0 – 2	0 – 4	0 – 4	1 – 10	3 - 19	1 – 12

Note. **$p < .01$. *$p < .05$. Concurrent partial correlations controlling for verbal ability and age are presented below the diagonal. Longitudinal partial correlations control for earlier verbal ability and concurrent age.

Table 3. *Descriptive Statistics and Correlations for Executive Function Tasks.*

		Wave 1				Wave 2		
		1	2	3	4	5	6	7
1	Tower of London	-	.26**	.27**	.26**	.27**	.32**	.37**
2	Bead Memory (Wave 1)	.12	-	.17*	.10	.42**	.33**	.25**
3	Day/Night Task (Pass/Fail)	.29**	.10	-	.24**	.06	.03	.16
4	Trucks Game (Pass/Fail)	.27**	.01	.22**	-	.09	.17*	.05
5	Bead Memory (Wave 2)	.15	.33**	.04	.01	-	.38**	.30**
6	Smiling Faces Efficiency (Test) (/s)	.25**	.17*	.03	.13	.19*	-	.35**
7	Arrows Task Efficiency (Test) (/s)	.31**	.16	.13	.05	.17*	.22**	-
	Mean	5.94	14.96	0.67	0.88	25.84	0.48	0.68
	SD	2.01	4.03	0.47	0.32	3.42	0.12	0.46
	Range	0 – 9	7 - 25	0 – 1	0 – 1	18 – 36	.16 - .79	0 – 1.86

Note. **$p < .01$. *$p < .05$. Concurrent partial correlations controlling for verbal ability and age are presented below the diagonal. Longitudinal partial correlations control for earlier verbal ability and concurrent age.

Table 4. *Concurrent and Longitudinal Correlations Between Theory of Mind, Executive Function, Verbal Ability and Social Competence.*

	Variable	1	2	3	4	5	6
1	ToM (Wave 1)	-	.44**	.32**	.47**	.35**	.34**
2	EF (Wave 1)	.32**	-	.16	.34**	.38**	.18^{+}
3	Teacher-Rated Social Competence (Wave 1)	.27**	.07	-	.17^{+}	.12	.31**
4	ToM (Wave 2)	.35**	.10	-.01	-	.41**	.29**
5	EF (Wave 2)	.18*	.25**	.05	.17*	-	.12
6	Teacher-Rated Social Competence (Wave 2)	.32**	.17^{+}	.29**	.32**	.14	-
	Concurrent Age	.19*	.36**	.17^{+}	.12	.20*	-.08
	Concurrent Verbal Ability	.53**	.36**	.15	.54**	.57**	.12
	Maternal Education	.25**	.28**	.05	.40**	.37**	.18^{+}
	Gender	-.05	-.08	.09	-.10	-.25**	.30**

Note. **$p < .01$. *$p < .05$. $^{+}p < .10$. Partial correlations are presented below the diagonal. Concurrent partial correlations control for verbal ability and age. Longitudinal partial correlations control for final age, earlier verbal ability and, for cross-lagged correlations, earlier scores on the dependent variable. *N* involving correlations for Teacher-Rated Social Competence at Wave 1 and Wave 2 = 103.

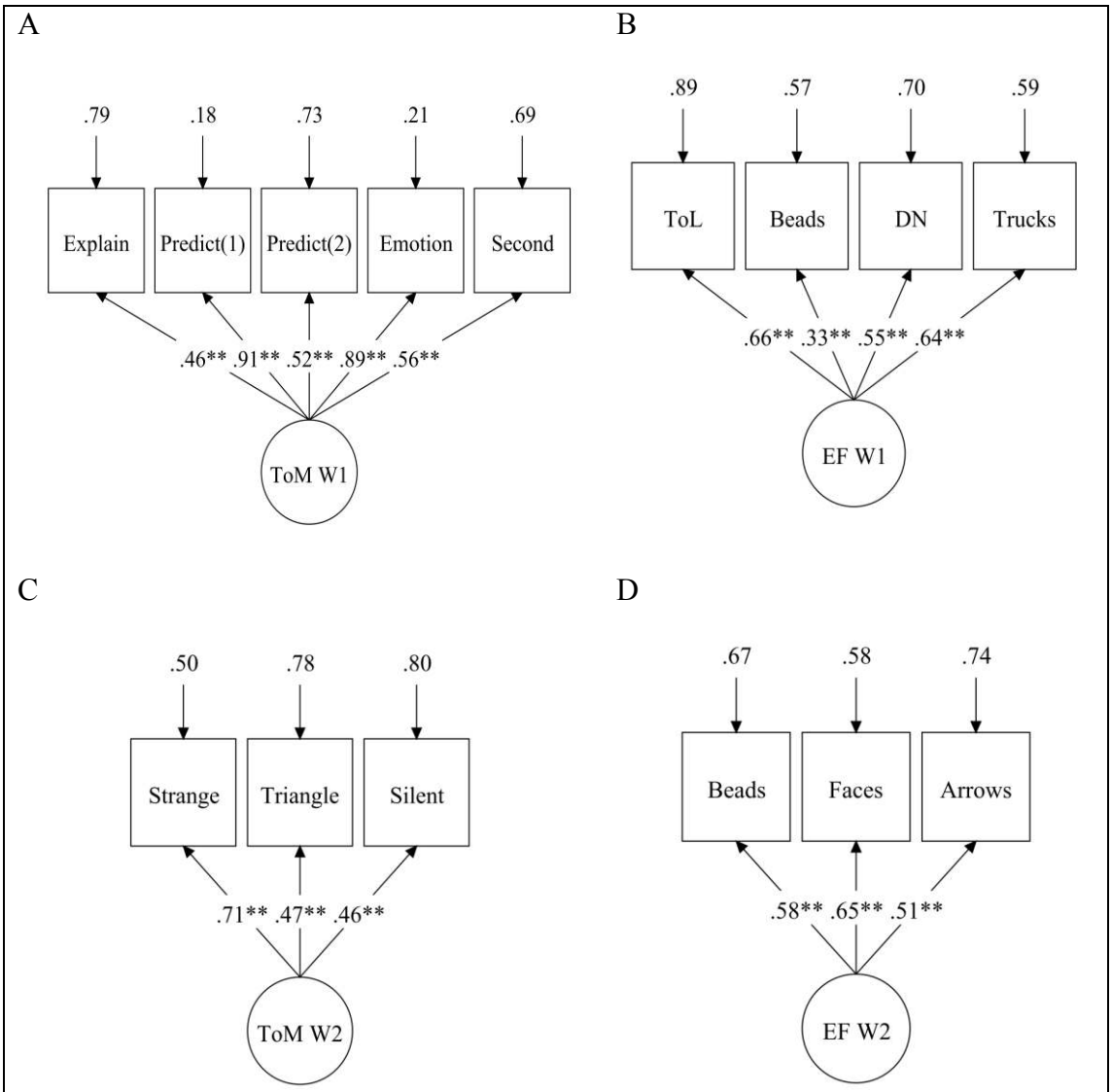

Figure 1. *Standardized Estimates for Measurement Models for ToM and EF Latent Factors at Wave 1 and Wave 2.*

Note. **p < .01. ToM = Theory of Mind. EF = Executive Function. W1 = Wave 1. W2 = Wave 2. Explain = Predict/Explain False Belief. Predict(1) = Predict False Belief I. Predict(2) = Predict False Belief II. Emotion = Infer/Explain Emotion False Belief. Second = Infer/Explain Second Order False Belief. ToL = Tower of London. DN = Day Night Task.

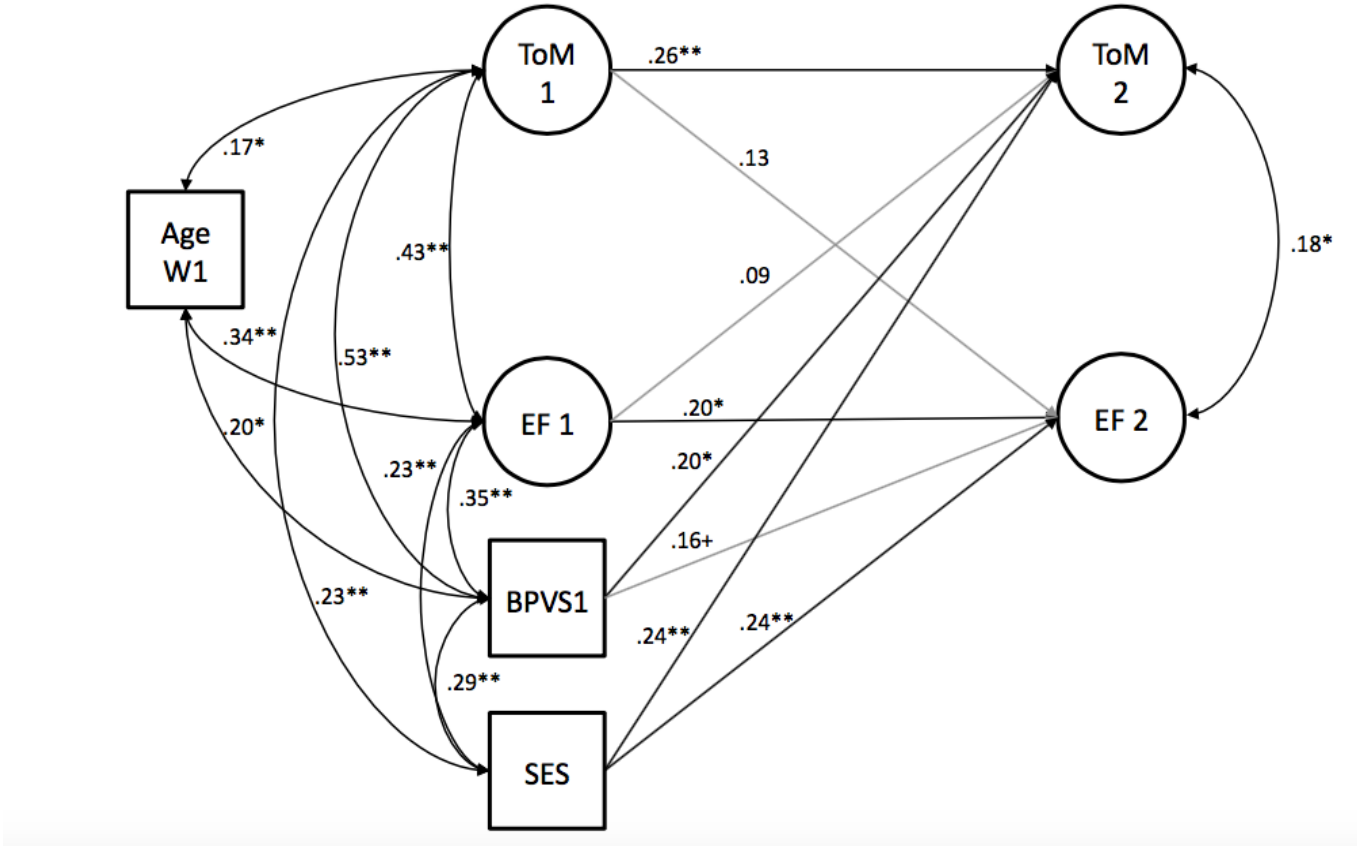

Figure 2. *Path Diagram Depicting Longitudinal Relations Between Theory of Mind and Executive Function*.
Note. $^{**}p < .01$. $^{*}p < .05$. $^{+}p < .10$. Paths with p > .05 are shown in grey. See Table S2 for unstandardized parameter estimates and standard errors for all parameters in the model. ToM = Theory of Mind, EF = Executive Function, BPVS = Verbal Ability, SES = Maternal Education, 1 = Wave 1, 2 = Wave 2.

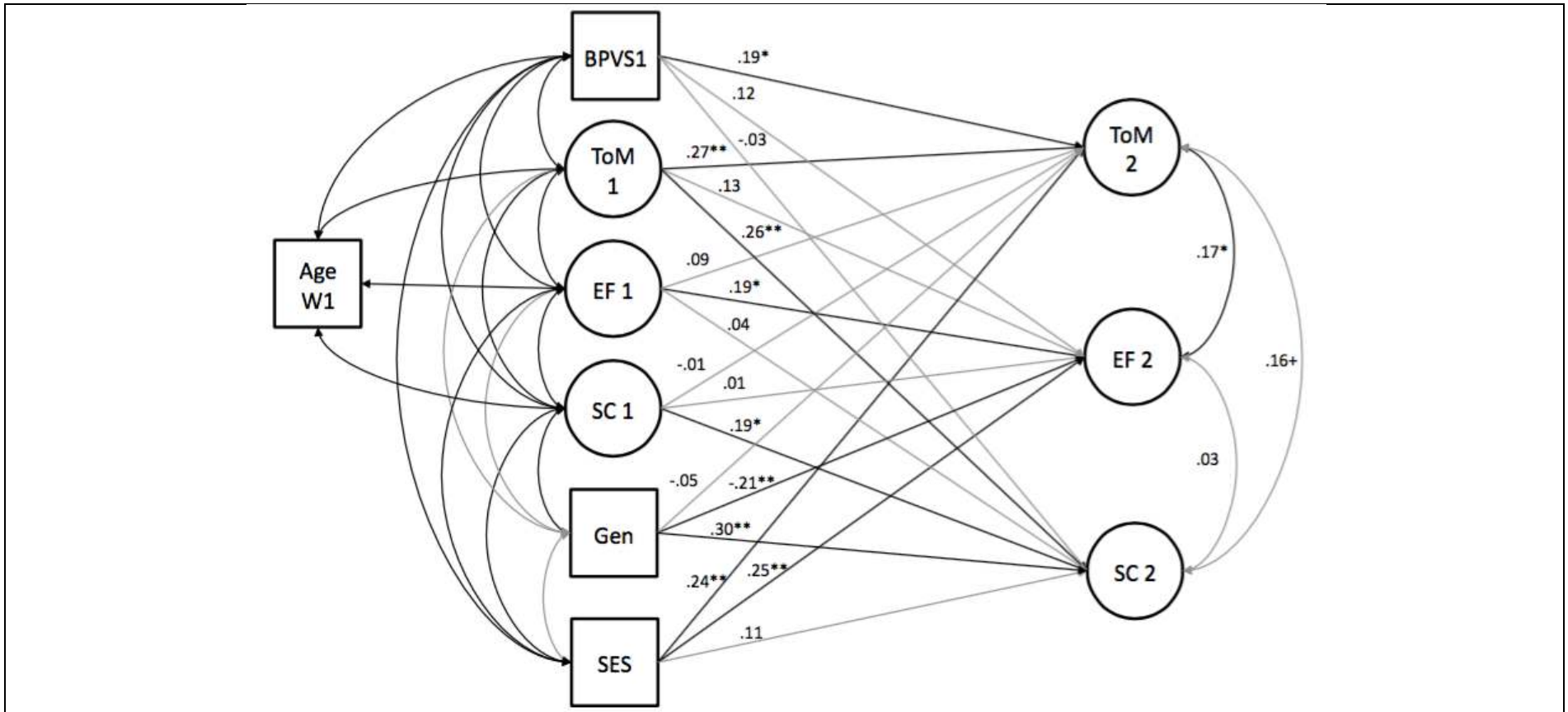

Figure 2. *Path Diagram Depicting Longitudinal Relations Between Theory of Mind, Executive Function and Social Competence*.
Note. $^{**}p < .01$. $^{*}p < .05$. $^{+}p < .10$. Paths with p > .05 are shown in grey. For clarity, covariances between predictor variables are not depicted. See Table S3 for unstandardized parameter estimates and standard errors for all parameters in the model. ToM = Theory of Mind, EF = Executive Function, SC = Teacher-Rated Social Competence, BPVS = Verbal Ability, SES = Maternal Education, 1 = Wave 1, 2 = Wave 2.

References

Apperly, I. A. (2011). *Mindreaders: The Cognitive Basis of "Theory of Mind."* Hove: Psychology Press.

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., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *Cognition*, *106*, 1093 - 1108.

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.

Astington, J. W. (2003). Sometimes necessary, never sufficient: False-belief understanding and social competence. In B. Repacholi & V. Slaughter (Eds.), *Individual Differences in Theory of Mind.* (pp. 13 - 38). Hove, UK: Psychology Press.

Asparouhov, T. & Muthèn, B.O. (2010a). Weighted least squares estimation with missing data. *Mplus Technical Appendix*. Los Angeles, CA: Muthèn and Muthèn.

Asparouhov, T. & Muthèn, B.O. (2010b). Plausible values for latent variables using *Mplus*. *Mplus Technical Appendix*. Los Angeles, CA: Muthèn and Muthèn.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False belief understanding in infants. *Trends in Cognitive Sciences*, *14*, 110-8.

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

Bartsch, K., & Estes, D. (1996). Individual differences in children's developing theory of mind and implications for metacognition. *Learning and Individual Differences*, *8*, 281-304.

Bartsch, K., & Wellman, H. M. (1989). Young children's attribution of action to beliefs and desires. *Child Development*, *60*, 946 - 964.

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

Byrne, B.A. (2012). *Structural Equation Modelling with Mplus.* New York: Routledge.

Carlson, S. M., Zelazo, P. D., & Faja, S. (2013). Executive Function. In P.D Zelazo (Ed.), *The Oxford Handbook of Developmental Psychology Volume 1: Body and Mind.* (pp. 706 - 743). Oxford, UK: Oxford University Press.

Carruthers, P. (1996). Autism as mindblindness: An elaboration and partial defence. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 257 - 273). Cambridge, UK: Cambridge University Press.

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.

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.

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*, 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.

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.

Dunn, J., Brown, J., Slomkowski, C., Tesla, C., & Youngblade, L. (1991). Young children's understanding of other people's feelings and beliefs: Individual differences and their antecedents. *Child Development*, *62*, 1352 - 1366.

Dunn, L., Dunn, D., Whetton, C., & Burley, J. (1997). *British Picture Vocabulary Scale* (2nd ed.). Windsor, UK: NFER Nelson.

Duval, C., Piolino, P., Bejanin, A., Eustache, F., & Desgranges, B. (2011). Age effects on different components of theory of mind. *Consciousness and Cognition*, *20*, 627 - 642.

Espy, K.A., Sheffield, T.D., Wiebe, S.A., Clark, C.A.C. & Moehr, M. (2011). Executive control and dimensions of problem behaviors in preschool children. *Journal of Child Psychology and Psychiatry, 52,* 33 – 46.

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3 1/2 - 7 years old on a Stroop-like day-night test. *Cognition*, *53*, 129-153.

Gresham, F. M., & Elliot, S. N. (1990). *Social Skills Rating System Manual*. Circle Pines, MN: American Guidance Service.

Happé, F. (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.

Harms, M.B., Zayas, V., Meltzoff, A.N. & Carlson, S.M. (2014). Stability of executive function and predictions to adaptive behavior from middle childhood to pre-adolescence. *Frontiers in Psychology, 5,* 331, 1 – 11.

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.

Hongwanishkul, D., Happaney, K.R., Lee, W.S. & Zelazo, P.D. (2005). Assessment of hot and cool executive function in young children: age-related changes and individual differences. *Developmental Neuropsychology, 28,* 617 – 644.

Hughes, C. (1998). Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind? *Developmental Psychology*, *34*, 1326-1339.

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., & 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* (7th ed., pp. 564 - 609). Hoboken, NJ: Wiley.

Hughes, C., & Ensor, R. (2005). Executive function and theory of mind in 2 year olds: A family affair? *Developmental Neuropsychology*, *28*, 645-668.

Hughes, C. & Ensor, R. (2007). Executive function and theory of mind: Predictive relations from ages 2 to 4. *Developmental Psychology, 43,* 1447 – 1459.

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., & Ensor, R. (2011). Individual differences in growth in executive function across the transition to school predict externalizing and internalizing behaviors and self-perceived academic success at 6 years of age. *Journal of Experimental Child Psychology*, *108*, 663-676.

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.

Huizinga, M., Dolan, C. V., & Van Der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia*, *44*, 2017 - 2036.

Jenkins, J. M., & Astington, J. W. (2000). Theory of mind and social behavior: Causal models tested in a longitudinal study. *Merrill-Palmer Quarterly*, *46*, 203 - 220.

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, 46*, 1417 – 1432.

Lagattuta, K.H., Sayfan, L. & Harvey, 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., & Banerjee, R. (2014). Promoting theory of mind in middle childhood: A training program. *Journal of Experimental Child Psychology*, *126,* 52 – 67.

McClelland, M.M., Cameron, C.E., Connor, C.M., Farris, C.L., Jewkes, A.M. & Morrison, F.J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary and math skills. *Developmental Psychology, 43,* 947 – 959.

Meins, E., Fernyhough, C., Johnson, F. & Lidstone, J. (2006). Mind-mindedness in children: Individual differences in internal-state talk in middle childhood. *British Journal of Developmental Psychology, 24,* 181 – 196.

Miyake, A. & Friedman, N.P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*, 8 – 14.

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.

Muthèn, B.O. (1997). Latent variable modelling with longitudinal and multilevel data. In A. Raftery (Ed.), *Sociological Methodology* (pp. 453 0 480). Boston: Blackwell.

Muthèn, B. O. (2004). *Mplus* Technical Appendices. Los Angeles, CA: Muthèn and Muthèn.

Muthèn, L. K., & Muthèn, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599 – 620.

Muthèn, L. K., & Muthèn, B. (2012). *MPlus: Statistical Analysis with Latent Variables User Guide.* (7th ed.). Los Angeles, CA: Muthèn and Muthèn.

Newton, E. & Jenvey, V. (2011). Play and theory of mind: Associations with social competence in young children. *Early Child Development and Care, 181,* 761 – 773.

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences*, *3*, 337 - 344.

Perner, J., & Lang, B. (2000). Theory of mind and executive function: Is there a developmental relationship? In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience* (2nd ed., pp. 150 - 181). Oxford: Oxford University Press.

Perner, J., Lang, B., & Kloo, D. (2002). Theory of mind and self-control: more than a common problem of inhibition. *Child Development*, *73*, 752-67.

Ratcliffe, M. (2007). *Rethinking Commonsense Psychology: A Critique of Folk Psychology, Theory of Mind and Simulation*. Basingstoke: Palgrave Macmillan.

Razza, R. & Blair, C. (2009). Associations among false belief understanding, executive function and social competence: A longitudinal analysis. *Journal of Applied Developmental Psychology, 30,* 310 – 320.

Rhoades, B. L., Greenberg, M. T., & Domitrovich, C. E. (2009). The contribution of inhibitory control to preschoolers' social–emotional competence. *Journal of Applied Developmental Psychology*, *30*, 310-320.

Ronald, A., Viding, E., Happé, F., & Plomin, R. (2006). Individual differences in theory of mind ability in middle childhood and links with verbal ability and autistic traits: A twin study. *Social Neuroscience*, *1*, 412-425.

Russell, J. (1996). *Agency: Its Role in Mental Development.* Hove, UK: Erlbaum Taylor & Francis.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *298*, 199 - 209.

Slaughter, V., & Repacholi, B. (2003). Individual differences in theory of mind: What are we investigating? In B. Repacholi & V. Slaughter (Eds.), *Individual Differences in Theory of Mind.* (pp. 1 - 12). Hove, UK: Psychology Press.

Stone, L. L., Otten, R., Engles, R., Vermulst, A. A., & Janssens, J. (2010). Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4- to 12-year-olds: A review. *Clinical Child and Family Psychology Review*, *13*, 254 - 274.

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

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scales*. Chicago: Riverside.

Tucker-Drob, E. M., & Briley, D. A. (2014). Continuity of genetic and environmental influences on cognition across the life span: A meta-analysis of longitudinal twin and adoption studies. *Psychological Bulletin*. doi: 10.1037/a0035893.

Watson, A. C., Nixon, C. L., Wilson, A., & Capage, L. (1999). Social interaction skills and theory of mind in young children. *Developmental Psychology*, *35*, 386-391.

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,* 655 – 684.

White, S., Hill, E., Happé, F. & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development, 80,* 1097 – 1117.

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.