Title: Minimal coherence among varied theory of mind measures in childhood and adulthood

Authors: Katherine Rice Warnell[1*] & Elizabeth Redcay[2]

[1] Department of Psychology, Texas State University, San Marcos, TX 78666
[2] Department of Psychology, University of Maryland, College Park, MD 20742
* Corresponding author

Correspondence:
Katherine Rice Warnell
Department of Psychology
Texas State University
San Marcos, TX 78666
Tel.: 512-245-5564
Email: warnell@txstate.edu

[Article: 6439 words; Abstract: 247 words; Figures: 1; Tables: 6; Supplemental Materials]

Funding: Internal support funds through the University of Maryland provided to ER and through Texas State University provided to KRW.

Acknowledgements: The authors wish to thank Seleste Braddock, Viviana Bauman, Robert Cai, Shannon Coveney, Callie de la Cerda, Laura Anderson Kirby, Sydney Maniscalco, Dustin Moraczewski, Jacqueline Thomas, Daniel O'Young, Kayla Velnoskey, Brieana Viscomi, & Marieke Visser for their assistance.

Keywords: theory of mind; early childhood; middle childhood; mentalizing; development; individual differences

Conflicts of Interest: None

Abstract

Theory of mind—or the understanding that others have mental states that can differ from one's own and reality—is currently measured across the lifespan by a wide array of tasks. These tasks vary across dimensions including modality, complexity, affective content, and whether responses are explicit or implicit. As a result, theoretical and meta-analytic work has begun to question whether such varied approaches to theory of mind should be categorized as capturing a single construct. To directly address the coherence of theory of mind, and to determine whether that coherence changes across development, we administered a diverse set of theory of mind measures to three different samples: preschoolers, school-aged children, and adults. All tasks showed wide variability in performance, indicating that children and adults often have inconsistent and partial mastery of theory of mind concepts. Further, for all ages studied, the selected theory of mind tasks showed minimal correlations with each other. That is, having high levels of theory of mind on one task did not predict performance on another task designed to measure the same underlying ability. In addition to showing the importance of more carefully designing and selecting theory of mind measures, these findings also suggest that understanding others' internal states may be a multidimensional process that interacts with other abilities, a process which may not occur in a single conceptual framework. Future research should systematically investigate task coherence via large-scale and longitudinal efforts to determine how we come to understand the minds of others.

1. Introduction

Although philosophers and psychologists have long been interested in how we think about other people’s thoughts (see Obiols & Barrios, 2009; Wellman, 2017 for historical review), Premack & Woodruff first introduced the term ‘theory of mind’ (ToM) in a 1978 paper that examined whether chimpanzees could infer human goals. The authors considered such mental state inferences to be evidence for a ToM—the capacity to represent the mental states of others. The term was quickly applied to human cognition research, and the subsequent 40 years have seen a rapid increase in articles investigating ToM across age groups and methodologies (for recent reviews, see Henry, Phillips, Ruffman, & Bailey, 2013; Mahy, Moses, & Pfeifer, 2014; Slaughter et al., 2015; Schaafsma et al., 2015; Schurz et al., 2014).

This wealth of ToM research has involved the creation of dozens of ToM measures, including tasks assessing false belief understanding (Wimmer & Perner, 1983), pragmatic language comprehension (Baron-Cohen et al., 1999; Happe et al., 1994; White et al., 2009), the ability to infer mental states from photographs of the eye region (Baron-Cohen et al., 2001), and reaction time when responding to actors’ beliefs (Apperly et al., 2011). Despite the surface differences between such measures, the field often treats all these social-cognitive paradigms as measures of ToM. As a result, any individual paper may select just one or two tasks in order to examine how ToM relates to another ability or differs between groups. However, theoretical proposals and recent reviews of neuroimaging and behavioral research suggest that ToM may not be a single construct (Apperly, 2012; Gerrans & Stone, 2008; Frith & Frith; 2008; Schurz et al., 2014; Schaafsma et al., 2015). In spite of these proposals, the extent to which varied ToM assessments relate to one another, and whether such measures do in fact capture a unitary construct, remains underexplored empirically.

The social cognitive literature contains long-standing theoretical discussions about the nature of ToM and its measurement. Much early work in this area was focused on the false belief task (e.g., Frith & Happe, 1994; Bloom & German, 2000), but more recent theoretical accounts have tackled the broader coherence of ToM. For example, Gerrans and Stone (2008) contrasted accounts of ToM as a domain-specific module versus accounts of ToM as multiple low-level domain-specific social processes intersecting with domain-general abilities including metarepresentation and executive function. Apperly (2012) similarly compared conceptual theories of ToM—which would argue for coherence among tasks—with cognitive theories, in which ToM is modelled not as a state of conceptual knowledge but as an interactive process spanning multiple cognitive abilities. Consistent with the latter perspective, Schaafsma and colleagues (2015) surveyed the vast array of different tasks measuring ToM and argued for the deconstruction of ToM into varied component processes (e.g., gaze processing, tracking intentions) rather than for ToM to be considered a single construct. In this framework, relations between ToM tasks could be due to common non-ToM demands (e.g., language, executive function) or due to common conceptual demands of specific types of ToM (e.g., false belief reasoning), rather than a broader conceptual coherence among all types of mental state reasoning.

In spite of this extensive theoretical discussion, empirical tests of ToM's unidimensionality have been limited. Papers introducing new ToM tasks often examine their relation with one or two existing tasks (e.g., Peterson & Slaughter, 2009; Beaumont et al., 2008; Devine & Hughes, 2013), but this literature may be biased to include positive relations, as new tasks that fail to show such relations may remain unpublished. Similarly, research comparing clinical and neurotypical groups on ToM batteries (e.g., Brent et al., 2004; Rosenblau, Kliemann, Heekeren, & Dziobek, 2015) does not directly comment on the underlying structure of ToM because group differences across tasks do not necessitate that performance on these tasks is correlated *within* subgroups. In the realm of neuroimaging research, meta-analytic evidence of overlapping activation across ToM tasks (e.g., Schurz et al., 2014; Molenberghs et al., 2016) does not necessarily indicate that such tasks tap into the same underlying mental process in particular individuals.

More targeted work has directly examined the relation between ToM measures in single samples. Some of the earliest work on this question examined relations between false belief measures in early childhood, finding, for example, that children who understood that others could have false beliefs about an object's location also understood that others could have false beliefs about an object’s appearance (e.g., Carlson & Moses, 2001; Hughes et al., 2000). This coherence among false belief measures in preschoolers is consistent with meta-analytic evidence that developmental trajectories of false belief acquisition are unaffected by task type (Liu, Wellman, Tardif, & Sabbagh, 2008; Wellman et al., 2001). More recently, researchers have also examined the relation between advanced theory of mind tasks at older ages. In middle childhood, there are significant correlations between children’s ability to answer explicit questions about mental states based on stories and their ability to answer similar questions based on video clips (Devine & Hughes, 2013; Devine & Hughes, 2016). Similarly, adults who are skilled at inferring complex emotional and mental states from pictures of the eyes show similar inferential skills when presented with pictures of the whole face and with spoken language (Meinhardt-Injac et al., 2018).

These existing studies of the coherence among ToM measures, however, are confounded by two important factors. First, such studies often use measures which assess conceptually-similar aspects of ToM (e.g., all false belief tasks or all tasks that involve explicitly inferring complex emotional states). Thus, coherence among tasks may be driven not by a common component underlying all mental state reasoning, but rather a conceptual commonality to one particular aspect of ToM. Second, the tasks used in existing studies often have very similar non-ToM cognitive demands (e.g., processing facial information). This confound means that such studies cannot address whether ToM represents a single construct as similar performance on these tasks may be due to the associated demands of other shared non-ToM component processes (Apperly, 2012; Gerrans & Stone, 2008). Thus, testing a wide array of diverse ToM measures would help establish whether ToM is a unitary construct.

A limited body of research has examined more diverse sets of ToM tasks within single samples and has produced inconclusive findings. For example, although some research has found that ToM tasks spanning modalities load onto a single factor in middle childhood (Osterhaus et al., 2016; Devine et al., 2016), other research has found evidence for much weaker patterns of relations on similar tasks in the same age range (Homer & Hayward, 2017; Rice et al., 2016). Further, even papers finding that one set of ToM tasks load onto a single factor have found that other ToM measures do not (Osterhaus et al., 2016; Devine et al., 2016), preventing conclusions about the coherence of ToM. Perhaps due to this lack of direct empirical research into the unidimensionality of ToM, a large number of studies continue to consider ToM a unitary construct (cf. Schaafsma et al., 2015), employing only one or two measures in order to capture ToM. Only by testing relations across tasks that assess different facets of ToM (e.g., false belief versus hidden emotions) and vary in their other non-ToM cognitive demands can we directly assess underlying ToM coherence (as opposed to coherence among other domains). This empirical exploration into the structure of ToM has both theoretical and practical relevance to the study of social cognition.

To behaviorally address the question of whether varied ToM measures form a unitary construct, we selected a range of widely-used ToM measures designed to capture individual differences in adult and child performance across a variety of specific tasks and modalities which have been argued to be important components of ToM (e.g., verbal versus non-verbal, affective versus cognitive, deliberate vs. automatic). The goal of this project was not to replicate literature examining the order of ToM concept acquisition (e.g., Wellman & Liu, 2004) or to determine if a narrow range of ToM tasks (e.g., affective verbal tasks or visual implicit tasks) were related to one another. We instead started with a broad slate of tasks, consistent with theoretical arguments that a diverse set of tasks might be the best route to understanding varied manifestations of ToM (Apperly, 2012); if these varied tasks did not cohere with each other, it would set the stage for future, more targeted work examining components of ToM. If, on the other hand, coherence emerged even on diverse tasks, such a finding would be strong evidence for unity in ToM.

We examined structure across these diverse ToM tasks in both children and adults, as the underlying structure of understanding others' thoughts may vary across development. Specifically, we administered multiple measures of ToM in early childhood (four-year-olds and six-year-olds), middle childhood (children aged 7-12), and adulthood. We selected a varied set of tasks for each age group, as older individuals are often at ceiling on measures (e.g., false belief tasks) appropriate for younger ages (Hughes, 2016; Lagattuta et al., 2015).

In our analysis of whether ToM measures were interrelated, several developmental patterns of results were possible. First, across all ages, different ToM measures could converge on a single factor. Second, children, but not adults, could show a single ToM factor. This would suggest there is a unitary mental inference ability early in development that becomes more task specific with age. Third, adults, but not children, could show convergence of ToM measures, potentially indicating that years of social experience crystallize ToM differences. In these scenarios, the middle childhood group could serve as an intermediary point between the preschoolers and adults. Finally, ToM might not form a unitary construct within any age group. Although conclusions from this study are necessarily limited to the specific set of tasks used, each of these potential findings has theoretical and practical implications for our understanding of ToM development, and can serve as a springboard for future research.

2. Methods

2.1. Participants

We initially collected data from 40 four-year-olds (14 males; average age 54 months), 38 six-year-olds (17 males; average age 79 months), and 40 children aged 7-12 (20 males, average age 10.09y). Our analyses suggested that ToM measures were not related to each other. To ensure that these results were not due to limited power, we then increased our sample size. Specifically, we targeted a sample size for each age group that would have 80% power to detect moderate correlations (approximately *r*=.35), an effect size consistent with studies that have examined developmental coherence among varied ToM and executive function measures (e.g., Carlson & Moses, 2001). To that end, we collected data from an additional 23 four-year-olds, 26 six-year-olds, and 26 school-aged children. Our final sample thus consisted of 63 four-year-olds (25 males; average age 54 months), 64 six-year-olds (29 males; average age 78 months), and 66 children aged 7 to 12 years (28 males, average age 9.82y). Children were recruited via a database of local families. All children were full-term, native English speakers, with no history of neurological damage, psychiatric disorders, head trauma, or psychological medications, and had no first-degree relatives with autism spectrum disorder or schizophrenia, as assessed via parent report.

Adult participants were recruited from the undergraduate student body of a large public university. The final adult sample was 222 adults (102 males) with an average age of 20.3 years (SD=3.0y). Adult participants were screened for neurological damage, for history of developmental disorders, and for first-degree relatives with autism spectrum disorder or schizophrenia through a self-report questionnaire.

2.2. Procedure

Children completed a behavioral battery consisting of ToM tasks (described below) and an IQ assessment. IQ was assessed with the Kaufman Brief Intelligence Test (KBIT-2; Kaufman & Kaufman, 2004), which yielded standardized scores for non-verbal, verbal, and full-scale IQ used in subsequent analyses. The order of tasks was identical across all participants within the same age group (i.e., four-year-olds, six-year-olds, middle childhood).

Additionally, for the early childhood sample, syntactic competence was assessed with the Sentence Structure subscale of the Clinical Evaluation of Language Fundamentals (CELF; Semel et al., 2003; Wiig et al., 2006). Four-year-olds and six-year-olds completed age-appropriate versions of the CELF (i.e., CELF-Preschool 2 and CELF-4). To match across these different assessments, age equivalencies were used in subsequent analyses.

Adults completed a ToM behavioral battery of five tasks (described below). As with the child sample, the order of the tasks was consistent across individuals. Given time constraints, IQ data was not collected for the adults.

2.3. ToM Assessments

2.3.1. General Task Selection

Across ages, we examined a broad range of tasks. Consistent with previous studies of ToM scaling, (e.g., Osterhaus et al., 2016; Hayward et al., 2017) we targeted measures which had high rates of adoption across various literatures, which were commonly discussed in these literatures as measuring ToM, and which produced individual differences. For each age group, we used tasks previously studied together in a single sample—in order to replicate and extend past results—and also employed widely-used instruments not previously examined concurrently with these other measures. For example, in four-year-olds, we selected multiple measures which had previously been shown to coalesce (false belief location, false belief contents, object appearance-reality; see next section for task specifics) as well as measures which had been attested to measure ToM but which appeared to assess different facets of ToM (e.g., understanding others' visual perspectives versus understanding faux pas) and had different non-ToM cognitive demands (e.g., processing visual versus verbal information). We also aimed, when possible, to use tasks that assessed similar underlying abilities across different age ranges (e.g., all groups completed an age-appropriate version of the Reading the Mind in the Eyes task; Baron-Cohen et al., 2001). In our six-year-old and middle childhood groups, we selected advanced ToM tasks which have been argued to cohere in a previous study, as well as measures argued to load onto a different factor (i.e., social reasoning versus understanding social transgressions; Osterhaus et al., 2016). For adults, task selection was more difficult given that many tasks used with children fail to produce variability in adults. In selecting adult tasks, we again aimed for a wide range of modalities and included some developmental commonalities when possible (e.g., a higher order theory of mind task as a corollary of first and second order belief understanding in children). Importantly, although there is ongoing debate in the field about whether and how some of our employed measures capture ToM (e.g., Reading the Mind in the Eyes; Oakley et al., 2016; Peterson and Miller, 2012), these instruments continue to be widely used in the literature as ToM measures, and thus their inclusion in the current battery has direct relevance to ongoing research programs. Finally, we note that there are dozens of ToM measures (e.g., Devine & Hughes, 2013; Dziobek et al., 2006; Keysar et al., 2003), with more under development, that we did not investigate in the current study. Our goal in task selection was to examine a wide range of modalities and task demands, consistent with the types of measures commonly used in the literature, in order to provide a starting point for empirical investigations of task coherence in ToM.

2.3.2. Early Childhood

We administered three tasks to both four- and six-year-olds: (1) a battery of traditional first- and second-order false belief tasks, (2) a preschooler-appropriate version of Reading the Mind in the Eyes Test where children had to make inferences about mental states from photos of the eye region (Simplified Eye Reading Test; Peterson & Slaughter, 2009; c.f. Baron-Cohen et al., 2001), and (3) an appearance-reality emotion task (Harris, Donnelly, Guz, & Pitt-Watson, 1986; Wellman & Liu, 2004) where children had to understand discrepancies between real and displayed emotion. Four-year-olds additionally completed an appearance-reality object task (Gopnik & Astington, 1988), in which they were presented with an object in which the external appearance did not match its true identity and evaluated both what they initially believed the object to be and what another child would think the object was. Six-year-olds also completed two additional tasks: (1) the Faux Pas Task (Baron-Cohen et al., 1999), in which they had to identify verbal faux pas from stories, and (2) the Restricted View Task (Lalonde & Chandler, 2002), in which they were shown a full picture which was then partially obscured and asked about what a character would think was depicted in the now ambiguous picture. For full details on the tasks used with the child samples, please see **Table 1** and **Supplemental Materials**.

2.3.3. Middle Childhood

We administered three tasks to children aged seven to twelve: (1) a school-age version of Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001), (2) the Strange Stories (Happe et al., 1994; White et al., 2009), in which children were presented vignettes containing mental states (e.g., white lie, double cross) and had to identify the motivation behind characters' statements, and (3) the Faux Pas Task (Baron-Cohen et al., 1999). Because the content of the Reading the Mind in the Eyes Test and the Faux Pas task differed between the early and middle childhood groups, we did not directly compare performance across these groups.

Table 1. Description of Tasks for the Early and Middle Childhood Theory of Mind Batteries.

Age Group	Task Name	Task Description
Early Childhood		
4 & 6-year-olds	False belief battery	-**False belief content**: a box contained an object different from that on the label, and children were asked what a character would think was in the box (2 trials presented) -**False belief location**: an object was moved unbeknownst to a character, and children were asked where the character would look (2 trials presented) -**Second-order false belief**: Children had to predict where a third character thought the protagonist would look for an object that was moved unbeknownst to the protagonist (2 trials presented)
4 & 6-year-olds	Simplified Reading the Mind in the Eyes Test	Children were presented with nine black-and-white photos of an adult's eye region and asked which of two emotions (e.g., serious vs. joking) best described the picture. This test is also referred to as the Simplified Eye Reading Test (SERT).

4 & 6-year-olds	Appearance-Reality Emotion	Children listened to five stories which the protagonist had reason to hide an emotional state and children had to identify the discrepancy between real and apparent emotion.
4-year-olds	Appearance-Reality Object	Children were shown an object that had a false appearance (e.g., chocolate that looked like a rock). After being shown the true identity, children were asked what they thought the object was when it was first presented, and were asked what a naïve character would think the object was. (2 trials presented)
6-year-olds	Faux Pas	Children listened to four short vignettes that each presented a social scenario and had to identify whether a faux pas was committed and, if so, why it was a faux pas.
6-year-olds	Restricted View	Children were first shown a simple line drawing of a common object (e.g., cow) and then the picture was mostly occluded, leaving a small portion—not identifiable as a cow—exposed. Children were then asked what two dolls who had not seen the whole picture would think it was. (2 trials presented)
Middle Childhood		
7- to 12-year-olds	School-age Reading the Mind in the Eyes Test	Children were presented with 28 black-and-white photos of an adult's eye region and asked which of four emotions best described the picture.
7- to 12-year-olds	Strange Stories	Children listened to eight stories involving mental states (e.g., double crossing, white lie) and were asked to explain the motivation behind a character's statement.
7- to 12-year-olds	Faux Pas	Children listened to eight short vignettes that each presented a social scenario and had to identify whether a faux pas was committed and, if so, why it was a faux pas.

2.3.4. Adults

Adult participants completed five tasks: (1) Spontaneous ToM Protocol (STOMP; Rice & Redcay, 2015), in which participants viewed two silent movie clips and described what happened in the scene; (2) Belief-Desires task (Apperly et al., 2011), a measure in which participants quickly answered questions about a character's true and false beliefs and desires; (3) pragmatic language comprehension (Koster-Hale, Dodell-Feder, & Saxe, unpublished), in which participants were presented with pairs of sentences and had to decide if one was an appropriate rejoinder to another; (4) the adult version of Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001); and (5) a higher-order theory of mind task (based on Kinderman et al., 1998, adapted in Rice & Redcay, 2015), in which participants listened to stories and answered ToM questions of increasing syntactic complexity. When possible, corrected scores for each task were calculated by adjusting for performance on a control task (e.g., adjusting higher-order ToM scores based on participant performance on control memory questions of equal syntactic complexity). In addition to varying on modality and affective content, the adult tasks also varied in whether they assessed more deliberate versus more rapid or spontaneous mentalizing. For example, the higher-order ToM stories and Reading the Mind in the Eyes explicitly asked participants to reason about mental states, whereas the STOMP measured the spontaneous tendency to mentalize. Due to difficulties with technical implementation, the Belief-Desires Task was added to the battery after data from the first set of adult participants was collected, so a smaller subset of participants completed all five tasks. For full details on the tasks and sample selection, please see **Table 2** and **Supplemental Materials**.

Table 2. Description of Tasks and Composite Scoring for the Adult Theory of Mind Battery.

Task Name	Task Description	Composite Scoring Procedure
Spontaneous Theory of Mind Protocol	Participants watched two silent film clips depicting socially-complex scenes and generated a spontaneous written description of the events in each clip.	The STOMP ratio was calculated by taking the number of internal statements and dividing by the number of total statements and multiplying by 100.
Belief Reasoning Speed	Participants were given information about a character's beliefs and desires and then asked which box the character would open, based on that information. For example, participants might be told that the red box had	Trials were collapsed together across desire type (positive or negative) Belief reasoning speed was calculated by subtracting reaction time for the true belief trials (where the character belief matched the real location of the

	yogurt, that the character liked yogurt and that he thought the green boxed contained yogurt.	food) from RT for the false belief trials (where the character's belief did not match the location of the food). This provided a measure of how much participants were slowed down by representing a false belief.
Pragmatic language comprehension	Participants were presented with pairs of sentences and had to determine if the pairs were a logical match. Sentences could match in the physical domain ("The highway is getting paved; The morning rush hour is starting even earlier than usual") or could match pragmatically ("I heard the new video game system just came out; We haven't seen our son in days"), including sarcasm. These pragmatic matches included sarcasm.	To control for baseline differences in verbal inferential ability, a composite score of pragmatic language ability was created, which subtracted out the percent accuracy score of the 44 physical causality items from percent accuracy on the 88 pragmatic items.
Reading the Mind in the Eyes Test	Participants were presented with a black-and-white photo of an adult's eye region and asked which of four emotions best described the picture.	Participants received one point for each correct mental state inference.
Higher-order ToM	Participants listened to a set of stories and evaluated whether statements about each story were true or false. Statements either were about mental states or factual events and varied in syntactic complexity from single clauses to up to four levels.	In order to capture higher-order mental state reasoning, analysis of the ToM stories was restricted to second and third order questions (45 memory and 45 ToM) and a composite score (Higher Order ToM) was calculated by subtracting percent accuracy on the memory items from percent accuracy on the ToM items.

2.4. Statistical analyses

For all age groups, we first examined the distribution of scores for each ToM measure, to test whether data produced robust variability and were not susceptible to ceiling or floor effects. For the early and middle childhood samples, given ordinal scoring and limited range of the child ToM assessments, relations between measures were analyzed using Spearman's rho rank-order correlation. As the ToM adult measures produced more continuous variability, adult correlations were initially analyzed using Pearson's r. We also conducted Bayesian analyses in order to quantify the strength of evidence in favor of the null versus alternative hypothesis (i.e., no relation between tasks vs. a relation between tasks).

After this initial examination, we used exploratory factor analysis to statistically examine underlying structure in the data in the adults only, given the relatively small sample sizes in the child groups (i.e., for each developmental sample, we had 80% power to detect correlations of .35). To determine whether such exploratory factor analysis reveals an underlying structure (i.e., how many factors to retain), researchers often rely on heuristics, such as visual examination of scree plots or retaining factors with an eigenvalue greater than one (Hayton, Allen, & Scarpello, 2004; O'Connor, 2000). Such heuristics, however, raise methodological concerns; retaining all eigenvalues greater than one may misestimate the number of components (Zwick & Velicer, 1986) and the examination of scree plots lacks reliability across users (Crawford & Koopman, 1979). As an alternative to these approaches, parallel analysis is one of the most accurate methods to determine underlying structure in data and we thus employed parallel analysis to conduct exploratory factor analysis in our current dataset (Glorfeld, 1995; O'Connor, 2000; Zwick & Velicer, 1986). Specifically, we generated one thousand random, normally distributed datasets that were similar to the original data in sample size and number of items simulated. Then, the eigenvalues from the real data set were compared to eigenvalues derived from the 95th percentile of the simulated data sets and we retained any components with eigenvalues greater than 95% of those generated by random simulation. All analyses were conducted with SPSS 24.0, with parallel analysis conducted using a macro from O'Connor (2000). Bayesian analyses of correlations were conducted using JASP 0.9.2.

3. Results

3.1. Early Childhood

3.1.1. Descriptive Statistics

All ToM tasks produced a wide range in performance (**Table 3**; see **Supplemental Materials** for histograms for all tasks across all age groups). Given that several tasks could only produce a limited range of values (e.g., integer scores from 0-4), non-parametric test were conducted on the data. Consistent with previous research, six-year-olds scored higher than four-year-olds on all tasks the groups had in common: the false belief index (Mann-Whitney *U* test=3426.0, p<.001), the appearance-reality emotion task (Mann-Whitney *U* test=3144.5, p<.001), and the Simplified Reading the Mind in the Eyes test (Mann-Whitney *U* test=2626.0, p<.01). There were no significant effects of gender for any tasks in the four-year-old group (ps>.05) and only the Faux Pas task showed an effect of gender in the six-year-old group (M_{Male}=2.1 items correct, M_{Female}=2.6 items correct, Mann-Whitney *U* test=2626.0, *p*=.021). Given that there was not a systematic effect of gender on performance, additional analyses collapsed across gender. We did repeat early childhood analyses including gender as a control variable, and there was no effect on our results (see **Supplemental Materials**).

On the CELF measure of syntactic comprehension, four-year-olds had an average age equivalency of 5.3 years (SD=1.0) and six-year-olds had an average age equivalency of 7.6 years (SD=1.1).

In order to determine if we needed to control for linguistic ability in subsequent analyses, we examined correlations with IQ and syntactic comprehension. We first collapsed across age groups and examined the three tasks completed by both four- and six-year-olds. Controlling for age in months, there was a significant relation between full-scale IQ and performance on the False Belief Composite (rho=.214, p=.02) and Appearance-Reality Emotion (rho=.295, p=.001), with stronger correlations with verbal IQ than non-verbal IQ for all three tasks. Even controlling for age, CELF age equivalency scores remained significantly associated with performance on all three of these tasks common to both age groups (rhos>.33, ps<.001). Next, we examined correlations for object appearance-reality task, which was completed only by the four-year-olds. Controlling for age in months, performance on this task was significantly correlated with CELF age equivalency (rho=.406, p=.001) and verbal IQ (rho=.162, p<.001). Finally, we analyzed the two tasks completed only by the six-year-old group, Faux Pas and Restricted view. Neither task was associated with IQ or language after controlling for age in months (rhos<.12). These results are consistent with evidence for a tight coupling between language ability and false belief and appearance-reality tasks (e.g., Milligan et al., 2007; Ruffman et al., 2003), with suggestions this coupling may be weaker in other ToM tasks (e.g., reasoning about social convention). Thus, in order to ensure that relations between ToM tasks were not driven by language ability, we examined partial correlations between tasks that accounted for common variance due to age, general verbal ability (verbal IQ), and syntactic competence (CELF age equivalency).

Table 3. Descriptive Statistics for the Early, Middle Childhood, and Adult Theory of Mind Batteries

	Mean	SD	Min.	Max.
Four-Year-Olds				
FB Index (% Accuracy)	43.4	35.5	0	100
Simplified Eyes (% Accuracy)	65.6	16.4	33.33	100
App-Reality Emo (% Accuracy)	25.7	26.3	0	100
App-Reality Object (% Accuracy)	71.0	30.5	0	100
Six-Year-Olds				
FB Index (% Accuracy)	89.1	19.5	0	100
Simplified Eyes (% Accuracy)	74.5	14.8	33.3	100
App-Reality Emo (% Accuracy)	63.8	36.3	0	100
Restricted View (% Accuracy)	57.2	22.6	0	100
Faux Pas (% Accuracy)	59.8	22.5	0	100
Middle Childhood				
School-age Eyes (% Accuracy)	65.6	11.7	42.9	85.7
Faux Pas (% Accuracy)	74.4	16.0	25.0	100
Strange Stories (% Accuracy)	74.0	13.4	37.5	100
Adult				
STOMP Ratio (% ToM statements)	31.1	9.9	0	56.0
Belief Speed (False belief – true belief, ms)	84	85	-160	350
False Belief Reasoning (ms)	756	162	250	1270
True Belief Reasoning (ms)	673	141	230	1110
Pragmatic (% ToM – % Control)	-3.6	6.8	-23.7	20.5
Pragmatic Inference (% Accuracy)	81.9	7.3	51.0	99.0
Physical Inference (% Accuracy)	85.5	7.7	50.0	100.0
Adult Eyes (% Accuracy)	71.4	11.3	38.9	97.2
Higher-Order (% ToM – % control))	-3.4	8.4	-36.1	20.5
Higher-order ToM Items (% Accuracy)	75.5	11.4	45.0	98.0
Higher-order Memory Items (% Accuracy)	78.8	10.9	45.0	98.0

3.1.2. Relations among ToM Tasks

Within the four-year-olds, only one significant relation between tasks emerged when controlling for age in months, syntactic competence (as measured by the CELF), and overall verbal ability (as measured by verbal IQ) (**Table 4**). Specifically, the false belief composite and the object appearance-reality task were significantly correlated. For the six-year-olds, there were no significant relations between any of the theory of mind tasks (*p*s>.1). Models that removed age did not change the pattern of results (see **Supplemental Materials**). The results in Table 4 were supported by Bayesian analyses, in which we calculated a Bayes Factor (BF_{10}) for each of the correlations. Unlike conventional null hypothesis significance testing, this process can determine the strength of evidence in favor of the null (i.e., that ToM tasks are not related; Wetzels & Wagenmakers, 2012). For all pairwise comparisons except between the false belief composite and object appearance-reality task, evidence was in favor of the null. Depending on the specific task comparison, the data were 2 to 6 times more likely to have occurred under the null (i.e., no relation) than the alternative (i.e., a relation; see **Supplemental Materials** for complete tables).

Within the composite false belief measure, individual subscales were correlated with each other. The three false belief tasks (location, contents, and second-order) were related to each other across both age groups, even after controlling for syntactic ability, overall verbal performance, and age (rhos > 0.36, ps < 0.001). Similarly, for four-year-olds, performance on the two object appearance-reality subscales (reasoning about their own previous belief and about a character's belief) were correlated with each other (uncorrected: rho = 0.346, p = .0055; corrected: rho = .238, p = .075).

Table 4. Relations among theory of mind tasks in early childhood

Four-year-olds				
	FB Index	Simplified Eyes	App-Reality Emo	App-Reality Object
FB Index	--	-.057	-.013	.274*
Simplified Eyes		--	-.148	-.187
App-Reality Emo			--	.004
App-Reality Object				---

Six-year-olds					
	FB Index	Simplified Eyes	App-Reality Emo	Restricted View	Faux Pas
FB Index	--	.029	.168	-.049	-.029
Simp. Eyes		--	.196	.182	-.205
App-Real Emo			--	.048	.135
Restricted View				---	-.060
Faux Pas					---

Note. Correlation values are Spearman's rho, controlling for age in months, verbal IQ, and age equivalency on the CELF Sentence Structure subscale. * *p*<.05

3.2. Middle Childhood

3.2.1. Descriptive Statistics

As with the early childhood measures, middle childhood ToM measures also produced a wide range of performance. Age was significantly positively related to all three tasks: faux pas performance (rho=.28, *p*=.037), Reading the Mind in the Eyes (rho=.25, *p*=.041), and Strange Stories (rho=.31, *p*=.011.). Controlling for age, full-scale IQ was significantly related to Strange Stories performance (rho=.39, *p*<.01) and marginally related to Reading the Mind in the

Eyes Test (rho=.24, *p*=.054), with both measures showing significant correlations with both verbal and non-verbal IQ. To be consistent with the early childhood analyses, we controlled for verbal IQ in the subsequent analyses by examining partial correlations. There were no significant differences between males and females on age, full-scale IQ, verbal IQ, or non-verbal IQ, or any of the three theory of mind tasks, so we collapsed across gender when examining relations between tasks.

3.2.2. Relations among ToM tasks

In a model correcting for age in months and verbal IQ, no significant relations emerged among the three theory of mind tasks (**Table 5**). We also tested models that controlled for non-verbal IQ and full-scale IQ instead of verbal IQ and the pattern of results was unaltered; no correlations were significant (*p*s>.1). Again, these results were supported by Bayesian analyses, which found the null to be more probable than the alternative for each covariate-corrected pairwise comparison.

Table 5. Relations among theory of mind tasks in middle childhood

	School-Age Eyes	Faux Pas	Strange Stories
School-Age Eyes	--	.164	.198
Faux Pas		--	.072
Strange Stories			--

Note. Correlation values are Spearman's rho controlling for age and verbal IQ.

3.3. Adulthood

3.3.1. Descriptive Statistics

All measures showed a wide range of scores, capturing individual differences (**Table 3**). Within the timed belief reasoning task, we replicated the finding of Apperly and colleagues (2011) that reasoning about false beliefs is slower than reasoning about true beliefs (on average, a difference of 84ms between conditions, *p*<.0001). Both the pragmatics and higher-order ToM tasks revealed an advantage in favor of non-ToM based inferences, with the average participant showing accuracy three percentage points higher on the non-ToM items (*p*s<.0001).

Pragmatic ability was significantly negatively correlated with age (*r*(198)=-.17, *p*=.017), but none of the other measures were associated with age. None of the measures showed a significant difference between males and females (*p*s>.05), although the female advantage on the Reading the Mind in the Eyes test approached significance (*t*(213)=-1.96, p=.051) and we thus controlled for age and gender in subsequent analyses.

3.3.2. Relation between Tasks

We first examined only participants with complete data on all five tasks (n=137). Pearson's correlations revealed no significant relations between any of the tasks and the correlations remained non-significant after controlling for age and gender (rs<.15; **Table 6**), with Bayesian analyses indicating that the null hypothesis was substantially more likely (i.e., over 3 times more likely; Jeffreys, 1961) for each of the pairwise comparisons. Examining correlations for the complete data set (i.e., for participants with data on at least two of the five tasks;

n=207) also did not indicate any significant correlations (*r*s<.15). Uncorrected correlations, as well as analyses that controlled for age but not gender, also failed to reveal any significant correlations between tasks (**Supplemental Materials**).

Table 6. Relations among theory of mind tasks in adulthood

	Spontaneous ToM	Belief Reasoning Speed	Pragmatics	Adult Eyes	Higher-Order ToM
Spontaneous ToM	--	-.023	.015	-.115	.125
Belief Reasoning Speed		--	.056	.115	.048
Pragmatics			--	.068	-.051
Adult Eyes				---	-.069
High-Order ToM					---

Note. Correlation values are Pearson's *r* controlling for age and gender.

Following this preliminary examination, we conducted an exploratory factor analysis on uncorrected data from the adult sample who completed all five tasks in order to determine whether the varied theory of mind tasks shared an underlying structure. Results from the parallel analysis suggest the data has a zero-factor solution (**Figure 1**). Specifically, none of the eigenvalues exceeded the 95th percentile cutoff from randomly generated data that mimicked the actual data in sample size and number of items.

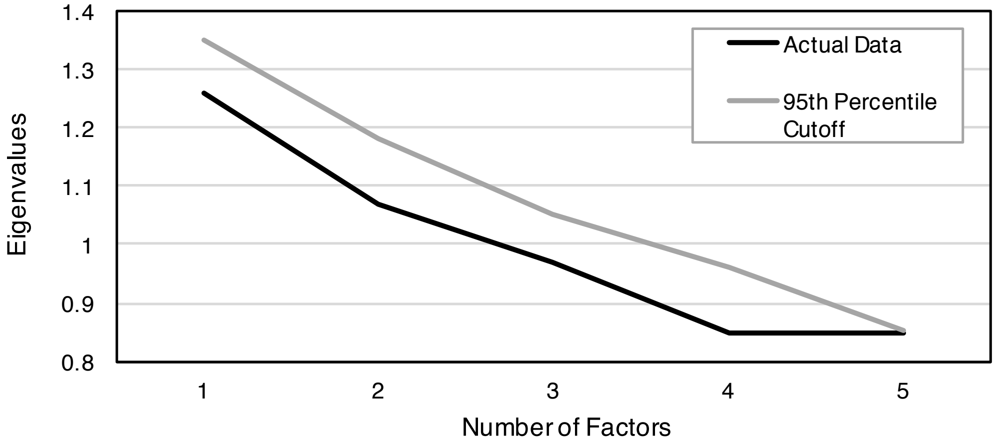

Figure 1. Parallel analysis examining structure of theory of mind in adulthood. One thousand random, normally distributed datasets similar to the original data in sample size and number of items were simulated. The 95th percentile of eigenvalues from this dataset were compared to the eigenvalues from the actual theory of mind tasks with results indicating a zero-factor solution.

4. Discussion

In the current study, we examined the relations between varied ToM measures at three time points across development. For the tasks used, no clear structure underlying ToM emerged for any developmental period. Specifically, after controlling for potential confounding variables (e.g., age, verbal ability), ToM tasks were minimally correlated in early childhood, in middle childhood, and in adulthood, a finding which was supported by Bayesian analysis that endorsed the null hypothesis. This finding could not be attributed to measurement issues such as dichotomous participant performance; all measures, even those typically conceptualized to represent all-or-nothing abilities, produced a range of scores. Instead, these results are consistent with past theoretical proposals (e.g., Apperly, 2012; Gerrans & Stone, 2008; Schaafsma et al., 2015) and suggest that ToM is a diverse construct that likely intersects with an array of other social and cognitive abilities, a finding with implications for both measurement and theory in the field of social cognition.

Consistent with past findings that conceptually-similar ToM tasks are related, we found positive relations in preschoolers between the three subscales that made up the false belief composite (two first-order content items, two first-order location items, and two second-order location items) and between the two subscales that made up the object appearance-reality composite (reasoning about one’s own previous beliefs and the beliefs of another). Additionally, the only significant positive relation found within any age groups was between the four-year-old false belief composite and the object appearance-reality composite. Similar correlations between false belief understanding and object appearance-reality tasks have been found in prior research, with effect sizes comparable to those in the current study (Gopnik & Astington, 1988). As some researchers, however, classify the two tasks as conceptually identical (Liu, Wellman, Tardif, & Sabbagh, 2008), these relations have limited bearing on the question of whether a unitary construct underlies all facets of ToM. Aside from the relation between false belief and object appearance-reality tasks, no clear relations between ToM tasks emerged in any of the three age groups.

Our findings suggest that ToM does not fractionate over development, but rather shows diverse structure from the preschool years. The continuity of this diverse structure, however, may be camouflaged by early childhood batteries which often heavily rely on false belief tasks that do converge (e.g., false belief location and false belief content tasks). These findings have implications for ToM research throughout the lifespan. Currently, many studies examining the effect of a particular circumstance, experimental manipulation, or intervention on ToM employ only a single ToM task. To give one highly-cited example, the claim that reading fiction improved theory of mind relied on the Reading the Mind in the Eyes Test (Kidd & Castano, 2013, 2019; but see Camerer et al., 2018). Further, many studies, especially with younger children, that do use multiple measures only use one item of each type (e.g., a single false belief location item and a single object appearance-reality item). All measures in the current study, however, even those typically conceptualized as dichotomous, showed a wide range of performance. Future work should ideally include several items from a variety of scales and be more precise about the exact facet of ToM interrogated by a particular measure.

The inclusion of a variety of different ToM tasks and other social cognitive and social perceptual measures in future studies will allow for a more precise understanding of the common and distinct correlates of different tasks, including relations with abilities such as basic biological motion perception (Miller & Saygin, 2013; Rice et al., 2016) and joint attention (Brooks & Meltzoff, 2015; Shaw et al., 2017; Sodian & Kristen-Antonow, 2015). As advocated by Schaafsma and colleagues (2015), a more detailed taxonomy of the individual basic level components of ToM assessments (e.g., perspective taking, emotion understanding, gaze following) will allow greater understanding of ToM. The authors draw a parallel to memory, in which such deconstruction has allowed for the identification of biologically-based component processes (e.g., short-term memory, long-term memory) and a more coherent examination of memory as a construct (Schaafsma et al., 2015).

The lack of relation among tasks in the current study is contrary to some existing research. For example, some studies in middle childhood have suggested that advanced theory of mind tasks converge on a single factor (Devein & Hughes, 2016; Osterhaus et al., 2016; but see Hayward et al., 2017 for findings that these tasks do not strongly cohere). This research, however, examined slightly different age ranges and cultural contexts than the current project. Future work should continue to use varied analytical approaches to examine coherence among ToM measures across cultures and ages to determine if there are other variables that affect the relative magnitude of task coherence. The psychometric properties of the particular ToM tasks used may also influence study results, as limited internal reliability across advanced ToM measures (e.g., Morrison et al., in press) may introduce measurement noise leading to inconsistent results across studies. Thus, we caution against over-interpreting any particular pairwise comparison of the present study and instead focus on our general pattern of results: ToM tasks do not coalesce as a single construct.

The finding that ToM measures do not converge may also seem to contradict well-established findings that ToM measures show systematic differences across ages and between clinical and typical groups, as well as findings that varied ToM measures activate similar neural networks. Group differences, however, do not statistically necessitate coherent individual differences (Gonzalez & Griffin, 2001; Hamaker et al., 2005; Na et al., 2010). General difficulties with social cognition may explain why children with and without autism show differential performance on ToM tasks. *Within* a particular group of children, however, what best predicts a child's relative performance on any one particular task may be due to the idiosyncratic demands of that instrument. Additionally, although a common neural network is implicated in ToM, more sophisticated neuroimaging analyses reveal more nuances in activation (Deen et al., 2015; Koster-Hale et al., 2017). Finally, it is possible that all ToM tasks do require common ToM conceptual knowledge, but as there is no meaningful variation in this basic ToM capacity across individuals, an individual differences approach will not capture this common component (cf. Apperly, 2012). Our results cannot directly speak to these hypotheses, but the finding that ToM fails to converge on an individual level extends, rather than contradicts, past research.

In the current study, we deliberately selected several types of measures that were likely to produce robust individual differences, rather than testing a specific a priori model of ToM (cf. Shaafsma et al., 2015). The tasks we used can be clustered on specific dimensions (e.g., Reading the Mind in the Eyes and appearance-reality emotion have affective information; Strange Stories and Faux

Pas tasks require reasoning about social narratives), but overall, they are quite dissimilar. Thus, this work is a stringent test of the hypothesis that all ToM tasks are related. Future studies should employ a targeted set of tasks in order to test specific underlying structures of social cognition. For example, one model could examine dissociations among tasks requiring rapid, implicit compared to deliberate, explicit inferences (cf. Apperly & Butterfill, 2009). Such targeted examinations may reveal stronger relations between similar tasks, although recent work does suggest that even similar implicit measures do not correlate (Kulke et al., 2018; Grosso et al., 2019; Poulin-Dubois & Yott, 2018; Powell et al., 2018).

In addition to limitations with task selection, another important limitation of the current study is that the sample size within each childhood age group was too small to allow for formal factor analysis and was powered to detect moderate effects. Future research should examine larger samples at each development time point. That said, if the relation between tasks is small enough to only be detected in very large samples, the coherence of ToM measures may have limited practicality. Future research should also follow children longitudinally, given evidence that early social abilities are predictive of later ToM (Brooks & Meltzoff, 2015; Sodian & Kristen-Antonow, 2015). Contrasting the state of extant theory of mind literature to the executive function literature may be instructive. Decades of research into executive function has included many studies with hundreds or even thousands of participants who were administered batteries designed to deconstruct component processes and identify their neural correlates (e.g., Carriedo et al., 2016; Gioia et al., 2002; Wiebe et al., 2008, 2011; Whelan et al., 2012), with evidence suggesting developmental fractionation in the components of EF (e.g., Brydges at el., 2014; Shing et al., 2010; Xu et al., 2013). Such a well-developed body of research does not yet exist for ToM and the current study was designed as a starting point spur future research and theoretical discussion.

Even if larger samples with more diverse measures continue to show a lack of underlying structure of ToM, this does not mean that ToM is an unimportant construct. For example, multiple ToM measures robustly capture between-group differences and age-related changes. More importantly, individuals clearly do use mental state understanding to navigate the social world. Rather than asserting, however, that this mental state understanding is a single representational ability, it may be more productive to consider the diverse ways in which understanding others' minds unfolds in the real-world (cf. Apperly, 2012). In some instances, we reason about emotions from visual information, in others, we parse verbal incidents, and in others, we take someone's visual perspective. The sophisticated understanding of others' minds that underscores mature human social cognition may be an emergent property of varied skills combined with certain social contexts. Critical examination of how and why we measure ToM will offer insight not just into existing ToM tasks but into cognition and behavior more broadly, as the lack of convergence among conventional ToM measures in the current study suggests that the best way forward in ToM research may be to take a step back.

References

Apperly, I. A. (2012). What is “theory of mind”? Concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology*, *65*(5), 825-839.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953.

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*(5), 1691-1703.

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407-418.

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

Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V., & Lawson, J. (2001). Are intuitive physics and intuitive psychology independent? A test with children with Asperger Syndrome. *Journal of Developmental and Learning Disorders*, *5*(1), 47-78.

Beaumont, R. B., & Sofronoff, K. (2008). A new computerised advanced theory of mind measure for children with Asperger syndrome: The ATOMIC. *Journal of Autism and Developmental Disorders*, *38*(2), 249-260.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25-B31.

Brent, E., Rios, P., Happé, F., & Charman, T. (2004). Performance of children with autism spectrum disorder on advanced theory of mind tasks. *Autism*, *8*(3), 283-299.

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology*, *130*, 67-78.

Brydges, C. R., Fox, A. M., Reid, C. L., & Anderson, M. (2014). The differentiation of executive functions in middle and late childhood: A longitudinal latent-variable analysis. *Intelligence*, *47*, 34-43.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child development*, *72*(4), 1032-1053.

Carriedo, N., Corral, A., Montoro, P. R., Herrero, L., & Rucián, M. (2016). Development of the updating executive function: From 7-year-olds to young adults. *Developmental Psychology*, *52*(4), 666.

Crawford, C. B., & Koopman, P. (1979). Note: Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills*, *49*(1), 223-226.

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, *25*(11), 4596-4609.

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. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the silent films and strange stories tasks. *Journal of Experimental Child Psychology*, *149*, 23-40.

Devine, R. T., White, N., Ensor, R., & Hughes, C. (2016). Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental Psychology*, *52*(5), 758.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., ... & Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of autism and developmental disorders*, *36*(5), 623-636.

Frith, U., & Happé, F. (1994). Autism: Beyond "theory of mind". *Cognition*, *50*(1-3), 115-132.

Frith, C. D., & Frith, U. (2008). Implicit and explicit processes in social cognition. *Neuron*, *60*(3), 503-510.

Gerrans, P., & Stone, V. E. (2008). Generous or parsimonious cognitive architecture? Cognitive neuroscience and theory of mind. *The British Journal for the Philosophy of Science*, *59*(2), 121-141.

Gioia, G. A., Isquith, P. K., Retzlaff, P. D., & Espy, K. A. (2002). Confirmatory factor analysis of the Behavior Rating Inventory of Executive Function (BRIEF) in a clinical sample. *Child Neuropsychology*, *8*(4), 249-257.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and psychological measurement*, *55*(3), 377-393.

Gonzalez, R., & Griffin, D. (2001). A statistical framework for modeling homogeneity and interdependence in groups. In G. J. O. Fletcher & M. S. Clark (Eds.), *Blackwell Handbook of Social Psychology.*

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, 26-37.

Grosso, S. S., Schuwerk, T., Kaltefleiter, L. J., & Sodian, B. (2019). 33-month-old children succeed in a false belief task with reduced processing demands: A replication of Setoh et al.(2016). *Infant Behavior and Development*, *54*, 151-155.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. (2005). Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis. *Multivariate Behavioral Research*, *40*(2), 207-233.

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129-154.

Harris, P. L., Donnelly, K., Guz, G. R., & Pitt-Watson, R. (1986). Children's understanding of the distinction between real and apparent emotion. *Child Development*, 895-909.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191-205.

Hayward, E. O., & Homer, B. D. (2017). Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. *British Journal of Developmental Psychology*, *35*(3), 454-462.

Henry, J. D., Phillips, L. H., Ruffman, T., & Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychology and Aging, 28*(3), 826-839.

Hughes, C. (2016). Theory of mind grows up: Reflections on new research on theory of mind in middle childhood and adolescence. *Journal of Experimental Child Psychology*, *149*, 1-5.

Hughes, C., Adlam, A., Happe, 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. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *41*(4), 483-490.

Jeffreys, H. (1961). *Theory of probability*. Oxford: UK Oxford University Press.

Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., ... & Kessler, J. (2010). Dissociating cognitive from affective theory of mind: a TMS study. *Cortex*, *46*(6), 769-780.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test*. John Wiley & Sons, Inc.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25-41.

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, *342*(6156), 377-380.

Kidd, D., & Castano, E. (2019). Reading Literary Fiction and Theory of Mind: Three Preregistered Replications and Extensions of Kidd and Castano (2013). *Social Psychological and Personality Science*, *10*, 522-531.

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, *89*(2), 191-204.,

Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage*, *161*, 9-18.

Koster-Hale, J., Dodell-Feder, D., & Saxe, R. (2012). [unpublished instrument]

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (in press). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development*.

Lagattuta, K. H., Kramer, H. J., Kennedy, K., Hjortsvang, K., Goldfarb, D., & Tashjian, S. (2015). Beyond Sally's missing marble: Further development in children's understanding of mind and emotion in middle childhood. *Advances in Child Development and Behavior*, *48*, 185-217.

Lalonde, C. E., & Chandler, M. J. (2002). Children's understanding of interpretation. *New Ideas in Psychology*, *20*(2), 163-198.

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology, 44*(2), 523-531.

Mahy, C. E., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience*, *9*, 68-81.

Meinhardt-Injac, B., Daum, M. M., Meinhardt, G., & Persike, M. (2018). The two-systems account of theory of mind: Testing the links to social-perceptual and cognitive abilities. *Frontiers in human neuroscience*, *12*, 25.

Miller, L. E., & Saygin, A. P. (2013). Individual differences in the perception of biological motion: links to social cognition and motor imagery. *Cognition*, *128*(2), 140-148.

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, *65*, 276-291.

Morrison, K. E., Pinkham, A. E., Kelsven, S., Ludwig, K., Penn, D. L., & Sasson, N. J. (in press). Psychometric evaluation of social cognitive measures for adults with autism. *Autism Research*.

Na, J., Grossmann, I., Varnum, M. E., Kitayama, S., Gonzalez, R., & Nisbett, R. E. (2010). Cultural differences are not always reducible to individual differences. *Proceedings of the National Academy of Sciences*, *107*(14), 6192-6197.

Oakley, B. F., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of abnormal psychology*, *125*(6), 818.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods*, *32*(3), 396-402.

Obiols, J. E., & Berrios, G. E. (2009). The historical roots of theory of mind: the work of James Mark Baldwin. *History of Psychiatry*, *20*(3), 377-392.

Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of advanced theory-of-mind tasks. *Child Development*, *87*(6), 1971-1991.

Peterson, C. C., & Slaughter, V. (2009). Theory of mind (ToM) in children with autism or typical development: Links between eye-reading and false belief understanding. *Research in Autism Spectrum Disorders*, *3*(2), 462-473.

Peterson, E., & Miller, S. (2012). The eyes test as a measure of individual differences: how much of the variance reflects verbal IQ?. *Frontiers in psychology*, *3*, 220.

Poulin-Dubois, D., & Yott, J. (2018). Probing the depth of infants' theory of mind: Disunity in performance across paradigms. *Developmental science*, *21*(4), e12600.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (in press). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*.

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

Rice, K., & Redcay, E. (2015). Spontaneous mentalizing captures variability in the cortical thickness of social brain regions. *Social Cognitive and Affective Neuroscience*, *10*(3), 327-334.

Rice, K., Anderson, L. C., Velnoskey, K., Thompson, J. C., & Redcay, E. (2016). Biological motion perception links diverse facets of theory of mind during middle childhood. *Journal of Experimental Child Psychology*, *146*, 238-246.

Rosenblau, G., Kliemann, D., Heekeren, H. R., & Dziobek, I. (2015). Approximating implicit and explicit mentalizing with two naturalistic video-based tasks in typical development and autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *45*(4), 953-965.

Sabbagh, M. A. (2004). Understanding orbitofrontal contributions to theory-of-mind reasoning: implications for autism. *Brain and Cognition*, *55*(1), 209-219.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*(2), 65-72.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9-34.

Semel, E., Wiig. E. H., Secord, W. A. (2003). Clinical Evaluation of Language Fundamentals (CELF – 4th Edition). San Antonio, TX: Harcourt Assessment.

Shaw, J. A., Bryant, L. K., Malle, B. F., Povinelli, D. J., & Pruett, J. R. (2017). The relationship between joint attention and theory of mind in neurotypical adults. *Consciousness and Cognition*, *51*, 268-278.

Shing, Y. L., Lindenberger, U., Diamond, A., Li, S. C., & Davidson, M. C. (2010). Memory maintenance and inhibitory control differentiate from early childhood to adolescence. *Developmental Neuropsychology*, *35*(6), 679-697.

Slaughter, V., Imuta, K., Peterson, C. C., & Henry, J. D. (2015). Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development*, *86*(4), 1159-1174.

Sodian, B., & Kristen-Antonow, S. (2015). Declarative joint attention as a foundation of theory of mind. *Developmental Psychology*, *51*(9), 1190.

Wellman, H. M. (2017). The Development of Theory of Mind: Historical Reflections. *Child Development Perspectives*, *11*(3), 207-214.

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 of theory-of-mind tasks. *Child development*, *75*(2), 523-541.

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic bulletin & review*, *19*(6), 1057-1064.

Whelan, R., Conrod, P. J., Poline, J. B., Lourdusamy, A., Banaschewski, T., Barker, G. J., ... & Fauth-Bühler, M. (2012). Adolescent impulsivity phenotypes characterized by distinct brain networks. *Nature Neuroscience*, *15*(6), 920.

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: revealing mentalizing impairments in autism. *Child Development*, *80*(4), 1097-1117.

Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology*, *44*(2), 575.

Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. A., Chevalier, N., & Espy, K. A. (2011). The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology*, *108*(3), 436-452.

Wiig, E., Secord,W., & Semel, E. (2006). *Clinical Evaluation of Language Fundamentals-Preschool 2nd Edition*. San Antonio, TX: The Psychological Corporation.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128.

Xu, F., Han, Y., Sabbagh, M. A., Wang, T., Ren, X., & Li, C. (2013). Developmental differences in the structure of executive function in middle childhood and adolescence. *PloS one*, *8*(10), e77770.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432.