6-30-2016

Do Adolescents with High Functioning Autism Verbally Mediate Theory of Mind?

Christina A. Irvine
University of Connecticut - Storrs, christina.irvine@uconn.edu

Do Adolescents with High Functioning Autism Verbally Mediate Theory of Mind?

Christina Anne Irvine, Ph.D.
University of Connecticut, 2016

Studies suggest that typically developing (TD) individuals verbally mediate theory of mind (ToM). Given that language deficits and ToM deficits are central to autism spectrum disorder (ASD), and given evidence that people with ASD may utilize visuospatial rather than verbal mediation, we sought to examine the role of language inner speech in ToM among adolescents with ASD. We utilized a false belief location-change dual task paradigm to examine whether a simultaneous verbal task, which inhibits inner speech, would differentially interfere with false belief task performance in participants with ASD. We predicted that results would indicate *less* reliance on inner speech among participants with ASD as compared to TD, and that language skills would be uniquely associated with ToM performance. Contrary to predictions and to the larger ToM literature, we found no group difference in false belief performance, and no additional decrement with verbal load. False belief performance was uniquely associated with VIQ within the ASD group, suggesting a critical role of language in ASD. This finding calls into question the theory that people with ASD "think in pictures" rather than words. Findings were consistent with a proposal that ToM involves both an implicit and an explicit system (Apperly & Butterfill, 2009): because people with ASD struggle with the former, they draw more heavily on an explicit ToM system to compensate for this deficit. Because verbal mediation is necessary for explicit cognitive tasks, such as explicit ToM, people with ASD may, in fact, rely on verbal mediation to “bootstrap” their ToM.

Keywords: autism spectrum disorder, verbal mediation, inner speech, theory of mind, dual task

Do Adolescents with High Functioning Autism Verbally Mediate Theory of Mind?

Christina Anne Irvine

B.A., Whitman College, 2008

M.A., University of Connecticut, 2014

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Do Adolescents with High Functioning Autism Verbally Mediate Theory of Mind?

Presented by

Christina Anne Irvine, B.A., M.A.

Major Advisor ____________________

Inge-Marie Eigsti, Ph.D.

Associate Advisor ____________________

Deborah Fein, Ph.D.

Associate Advisor ____________________

Marie Coppola, Ph.D.

University of Connecticut
2016

Acknowledgments

First, I am truly grateful to the participants and families who generously volunteered their time to participate in this study. This project would not have been possible without the support and encouragement of my many research and clinical mentors. Thank you to my advisor and research mentor, Inge-Marie Eigsti, for wholeheartedly fostering my growth as a researcher. Her intellect, curiosity, and enthusiasm are an inspiration to me. Thank you also to Deborah Fein, for her generosity and encouragement over the years. I would also like to thank Marie Coppola, for her kindness and dedication of time to this project. I would to express my sincere thanks to my readers: Stephanie Milan, for her generous assistance, and Marianne Barton, whose mentorship and support have profoundly influenced my professional and personal development.

In addition to my committee, I would like to thank many others with whom I have had the pleasure of working with at UConn. Thank you to Dev Dalal for consultation on data analysis. Additional thanks to our research assistants: Molly Waterman, Paris Shah, Morgan Smith, Ryan Morris and Bonnie Stas, for their help piloting and running participants for study 2 of this project. A special thanks to Allison Fitch and Christopher Andrade for their performances as Annie and the Custodian. I would also like to express my gratitude for the graduate students of the Developmental Cognitive Neuroscience team: Allison Canfield, Brian Castelluccio, Joshua Green, and Anders Hogstrom for their constant good humor and moral support, and a special thanks to Jessica Mayo, for her partnership in recruitment, data collection, data management, and many hours logged on the road. Many thanks also to my cohort for their encouragement over the years. And of course, thank you to my family, especially Matteo Leggett, who constantly supported and grounded me throughout this process, and with whom I share this achievement. Finally, I dedicate this dissertation in memory of my mom, Jill Irvine. Her example and love continue to guide me every day.

Table of Contents

Do adolescents with high-functioning autism verbally mediate theory of mind?

The relationship between language and thought is an ongoing topic of debate among psycholinguists. Inner speech, or “intrapersonal communication,” is defined as thought in the form of language that is internal to oneself. According to Vygotsky (1987), linguistic thought is the primary medium for representing and relating abstract concepts and the foundation for flexible cognition and self-regulation. Vygotsky (1987) argued that linguistically-mediated interpersonal exchanges with caregivers early in life serve to regulate the child’s emotions and behavior; over time, these dialogues become more intrapersonal, and the child is able to regulate herself through overt self-talk or “private speech” (Winsler & Naglieri, 2003). By middle-childhood, this self-talk becomes internalized as inner speech.

In children with ASD, however, social and communication deficits are present from early in development. By the Vygotskian (1987) account that internalization of speech is scaffolded by social dialogue, people with ASD may demonstrate diminished private, or “inner” speech (Fernyhough, 1996). Given the importance of inner speech for higher-order mental processes, such as theorizing about the minds of others, as reviewed below, and given the prevalence of theory of mind (ToM) deficits among people with ASD, we were interested in exploring whether people with ASD verbally mediate ToM. In the current study, we examined the role of inner speech in ToM reasoning among adolescents with high-functioning ASD and typical development (TD).

Inner speech supports higher-order thought. Several studies have demonstrated that inner speech mediates higher-order thought and executive control (Barkley, 1997). A common way of assessing the contribution of inner speech to the execution of cognitive tasks is to

examine task performance when inner speech is suppressed, often by increasing verbal load via articulation of a repeated word or phrase. Such studies find that inner speech is selectively disrupted, as shown by reduced recall of verbally-encoded information while repeating a word aloud (Murray, 1967), while visuospatial encoding is preserved, as shown by intact spatial-sequence recall during articulatory suppression (Vandierendonck, Kemps, Fastame, & Szmalec, 2004). If a cognitive task requires verbal mediation, performing the task under a condition of verbal load hinders task performance.

Using this *dual task* methodology, Baddeley, Chincotta, and Adlam (2001) found that verbal load selectively interferes with the ability to alternate between mental operations such as addition and subtraction, suggesting that inner speech mediates task switching. Studies have also found that inner speech suppression selectively interferes with tower task performance (Lidstone, Meins, & Fernyhough, 2010) and that, in young children, private speech use correlates with tower task performance (Fernyhough & Fradley, 2005), suggesting that inner speech also mediates the executive function of planning. Together, these findings suggest that language serves as a primary medium through which people internally inhibit and plan behavior, and hold in mind and link together competing representations or concepts (Carruthers, 2002).

Language and theory of mind. One type of higher-order thought central to the human experience is our ability to theorize about the minds of others. Theory of mind (ToM) involves understanding that other people have ideas and beliefs that are distinct from one's own, and thus allows for the attribution of mental states, such as thoughts, feelings, desires, and beliefs, to others. False belief reasoning, or the ability to reason about another person's beliefs regardless of the veracity of these beliefs, is often taken as a proxy for the broader construct of ToM. A variety of experimental tasks are used to test belief reasoning, including location-change false belief

tasks and appearance-reality tasks. Both of these tasks involve an explicit dissociation of reality from another person's (incorrect) belief, as well as the ability to inhibit the default "true" response in favor of the "false belief" response. Belief reasoning emerges in early childhood; typically developing children begin to pass false belief tasks at age four (Wellman, 1990).

Representing another person's mental state requires holding in mind multiple representations and involves executive demands, such as planning (Carlson, Mandell, & Williams, 2004) and inhibitory control (Bialystok & Viswanathan, 2009). Several authors have suggested that the development of executive functions is integral to the development of ToM (Carlson et al., 2004). Therefore, language may offer a unique medium not only for executive functions, but also for representing the mental states of others. Deaf children who learn sign language at an early age perform better on false belief tasks than late sign language learners (de Villiers & de Villiers, 2005; Woolfe, Want, & Siegal, 2002). Moreover, a meta-analysis found a significant relationship between language skills and false belief competence independent of age among children younger than age 7, with language ability accounting for 18% of the variance (Milligan, Astington, & Dack, 2007).

There are, in fact, specific linguistic structures that may be critical for representing mental states. Sentences with a primary clause and a complement clause with a belief verb (such as "think," "assume," and "hope") serve to represent the mental state of another person. For example, in the sentence "Sally thinks that Bob is sleeping," [Bob is sleeping] is the primary clause, and [Sally thinks that] is the complement clause. The structure of this sentence provides simultaneous representation of two distinct realities: Sally thinks that Bob is sleeping although Bob may be awake (de Villiers, 2007).

Inner speech and theory of mind. To directly test the hypothesis that inner speech mediates belief reasoning, Newton and de Villiers (2007) conducted a study presenting a belief-reasoning task while inner speech was or was not suppressed. The study used a between-subject design; half of the participants completed a computerized belief-reasoning task (half of whom completed a false belief task in which a puppet places an object under a box and leaves, another puppet enters the room and moves the object to one of two other boxes, and the first puppet re-enters the room; the other half of whom completed a true belief task, which was the same as the above mentioned task except that the first puppet remains in the room when the object is moved). While completing the belief-reasoning task, participants performed either a verbal load task (shadowing a complex narrative) to suppress their inner speech. The other participants performed either the false or true belief task along with a nonverbal load task (tapping in rhythms), presumably imposing a similar cognitive load, but without disrupting inner speech. False belief accuracy was high for the nonverbal load group, but at chance for the verbal load group, suggesting that inner speech uniquely mediates ToM.

Theory of mind in ASD. Deficits and delays in ToM are considered a central feature in ASD. Whereas typically developing children begin to pass belief-reasoning tasks at around four years of age (Wellman, 1990), children with ASD do not pass these tasks until age nine, on average (Happé & Frith, 1996). For children with ASD, the severity of social deficits is associated with belief-reasoning deficits (Frith, Happé, & Siddons, 1994). Interestingly, children with ASD use belief language less frequently than their typically developing peers (Lind & Bowler, 2009), and production of sentence complements among children with ASD is associated with belief reasoning ability (Hale & Tager-Flusberg, 2003).

Inner speech and theory of mind in ASD. Several studies have demonstrated that people with ASD rely less on inner speech to mediate and mentally represent information. For example, evidence from fMRI studies suggests that people with ASD rely more heavily on visuospatial than on linguistic strategies when thinking about abstract information (Kana, Keller, Cherkassky, Minshew, & Just, 2006). Several studies have found that suppressing inner speech in ASD causes *less* performance decrement during set-shifting (Whitehouse, Maybery, & Durkin, 2006) or planning (Wallace, Silvers, Martin, & Kenworthy, 2009; Williams, Bowler, & Jarrold, 2012), suggesting less verbal mediation of executive tasks among people with ASD. Engaging in FB tasks requires top down executive processing, including set shifting, planning, and inhibitory control. Individuals with ASD may rely less on inner speech to mediate both executive and ToM tasks.

It appears that individuals with ASD demonstrate less verbal mediation of executive functions. However, findings do indicate that inner speech suppression affects short-term memory in ASD (Williams et al., 2012), suggesting that, unlike executive functions, people with ASD verbally mediate short-term memory. Fernyhough (2008) proposed an interesting explanation for this discrepancy: whereas *monologic* inner speech affords linear processes, such as short-term and working memory (e.g. the phonological loop), *dialogic* inner speech involves an internal “conversation” or “dialogue” between different perspectives held by oneself, thus affording cognitive flexibility and executive control. In other words, monologic inner speech can be described as ‘for oneself,’ whereas dialogic inner speech can be described as ‘with oneself’ (Williams et al., 2012). Consistent with Vygotsky’s (1987) theory of a developmental internalization of language, and because ASD involves deficits engaging in dialogue with *others*, people with ASD may also have deficits “engaging in dialogue with themselves.” Because ToM

inherently involves simultaneous representation of multiple perspectives, ToM likely involves dialogic inner speech. Deficits in dialogic inner speech may therefore contribute to the ToM deficits seen in ASD.

Despite delays in ToM development, many children with ASD ultimately pass belief-reasoning tasks. Therefore, although people with ASD have more difficulty with ToM than their peers, they do not categorically *lack* theory of mind. Instead, it appears that the ability to reason about others' minds is a gradated skill that varies in different contexts and under different constraints. For example, people with ASD tend to demonstrate stronger ToM abilities in experimental settings as compared to naturalistic settings (Apperly & Butterfill, 2009). Given that people with ASD, particularly those with intact cognitive skills, are eventually able to demonstrate ToM, though with varying efficiency (Channon, Charman, Heap, Crawford, & Rios, 2001; Kaland, Callesen, Møller-Nielsen, Mortensen, & Smith, 2008), we sought to explore the role of inner speech in thinking about the minds of others in individuals with high-functioning ASD. In the present study, we examined the role of inner speech in ToM (using a false belief task) in adolescents with ASD.

Current study. Given evidence of global ToM deficits in ASD, and given evidence of inner speech deficits in people with ASD (Wallace et al., 2009; Whitehouse et al., 2006; Williams et al., 2012), we hypothesized that participants with ASD verbally mediate ToM less than their TD peers. To test this hypothesis, in Study 1, we examined the degree to which a verbal load dual task, which suppresses inner speech, interfered with false belief task performance in adolescents with high-functioning ASD as compared to TD adolescents. Participants with ASD and TD completed true and false belief tasks while engaging in a verbal interference dual task, designed to suppress inner speech, and a spatial load dual task, designed

to leave inner speech intact. We measured task performance in terms of accuracy. We also measured reaction time (RT) for each trial, as we anticipated that this metric might be more sensitive to group differences. We anticipated that verbal mediation would be reflected in a greater performance decrement (reduced accuracy or greater RT) under verbal load as compared to spatial load. Therefore, we predicted a performance decrement for TD participants on false belief trials under verbal load, but an absence of this task by condition interaction in the ASD group.

We were also interested in exploring the relationship between ToM, external language skills, and EF. Considering evidence that language is integral to false belief in TD (Milligan et al., 2007) but appears to be underutilized during mediation of cognitive tasks among people with ASD, we predicted a significant correlation between language skills and false belief performance among TD participants, but a minimal correlation between language skills and false belief performance among ASD participants. Moreover, given studies suggesting a relationship between false belief and inhibitory control (Bloom & German, 2000; Leslie, Friedman, & German, 2004) as well as executive functions more generally (Bialystok & Viswanathan, 2009), we anticipated that direct and indirect measures of inhibitory control and general executive functions EF scores would correlate with false belief performance for all participants.

In Study 2, we collected data to test the comparability of cognitive load imposed by the verbal load versus spatial load task in order to confirm that any load-driven performance differences in Study 1 were due to the verbal versus spatial nature of the dual tasks rather than task demands more generally.

Study 1: Belief-Reasoning Dual Task

Methods

Participants. Participants were 32 adolescents between the ages of 12 and 18 with autism spectrum disorder (ASD; n = 15) or typical development (TD; n = 17). The groups did not differ on gender, age, full scale IQ (FSIQ), or general language ability. There were significant group differences, however, in both verbal IQ (VIQ) and nonverbal IQ (NVIQ): the TD group VIQ was 1 SD higher than the ASD group, and the ASD group NVIQ was 1 SD higher than the TD group. Participant characterization data are shown in Table 1.

Table 1. *Characteristics of Autism Spectrum Disorder (ASD) and Typically Developing (TD) groups.*

	ASD	TD	F/χ^2	p
n	15	17		
Sex (M:F)	13:2	14:3	$\chi^2 = 0.11$	0.74
Age	14.1 (1.6)	14.4 (1.9)	0.25	0.62
	12-18	12-17		
FSIQ	102.4 (9.7)	103.4 (8.6)	0.09	0.77
	85-121	88-124		
VIQ	9.7 (2.7)	11.5 (2.0)	4.89	0.04
	5-13	9-16		
NVIQ	11.2 (2.4)	9.5 (1.9)	4.68	0.04
	7-15	6-13		
SCQ (Raw score)	27.2 (13.4)	2.4 (2.1)	56.72	0.001
	16-59	0-7		
CELF Core Lang	108.6 (10.3)	115.0 (9.7)	3.12	0.09
	93-123	93-129		
D-KEFS C-W Inhib	10.3 (1.4)	11.8 (2.6)	3.89	0.06
	8-13	5-16		
BRIEF GEC	62.2 (7.9)	48.5 (8.0)	22.29	0.001
	44-78	38-61		

Table data are reported as mean (*SD*), range. All measures are given as standard scores, unless otherwise noted. FSIQ = Stanford-Binet Intelligence Scales, Fifth Edition, Full Scale IQ; VIQ = Stanford-Binet Verbal IQ; NVIQ = Stanford-Binet Nonverbal IQ; SCQ = Social Communication Questionnaire; CELF = Clinical Evaluation of Language Fundamentals; D-KEFS = Delis-Kaplan Executive Function System; BRIEF = Behavior Rating Inventory of Executive Function.

All participants had full-scale IQ scores in the average range (SS > 85) as measured by the Stanford-Binet Intelligence Scales, Fifth Edition, abbreviated (Roid, 2003). Inclusion criteria for each group were as follows:

ASD group. Participants met diagnostic criteria based on DSM-IV-TR criteria (American Psychiatric Association, 2000), confirmed using the Autism Diagnostic Observation Schedule (ADOS; Lord, Rutter, DiLavore, & Risi, 2002), the Social Communication Questionnaire (SCQ; Rutter, Bailey, & Lord, 2003), and clinical judgment. All participants had an ADOS composite score at or above the clinical cutoff of 7 and an SCQ score at or above the clinical cutoff of 15.

TD group. Participants in the TD group did not meet criteria for ASD based on clinical judgment or the SCQ. Participants had no first-degree relatives with an ASD diagnosis. In order to avoid a hyper-normative group, however, TD children were not excluded for other learning or psychiatric disorders.

Exclusion criteria. Participants were excluded from all groups if they exhibited symptoms of major psychopathology that would impede study participation, a history of traumatic brain injury with loss of consciousness, or other neurologic disorders or known genetic syndromes.

Participants were part of a larger study on social cognition in ASD. All procedures were approved by University of Connecticut Institutional Review Board. Parents provided written assent and participants provided written assent for study participation; participants received financial remuneration.

Measures. Participants completed a comprehensive assessment of autism symptoms, verbal IQ, nonverbal IQ, working memory, and executive functions. ASD diagnosis was

confirmed for the ASD group using the *Autism Diagnostic Observation Schedule* (ADOS; Lord et al., 2002), the "gold-standard" measure for assessing and diagnosing autism spectrum disorder. Autism symptom severity was measured using the *Social Communication Questionnaire, Lifetime Version* (SCQ; Rutter et al., 2003), a parent-report measure of lifespan autism symptoms.

FSIQ was measured using the *Stanford-Binet* (Roid, 2003), with a Vocabulary (verbal routing) subtest to assess verbal IQ (VIQ), and Nonverbal Reasoning (nonverbal routing) subtest to assess nonverbal IQ (NVIQ). General language ability was measured using the *Clinical Evaluation of Language Fundamentals* (CELF; Semel, Wiig, & Secord, 2003) Core Language composite index.

To measure EF, participants completed the Color-Word Interference: Inhibition subtest from the *Delis-Kaplan Executive Function System* (D-KEFS; Delis, Kaplan, & Kramer, 2001), which measures inhibitory control. Parents completed the *Behavior Rating Inventory of Executive Function* (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000), a parent-report measure of executive functions, which generates several specific domain scores as well as a Global Executive Composite index, with higher values indicating greater executive dysfunction.

Experimental task. The experimental belief-reasoning dual task was designed to test false belief comprehension under dual task conditions. Similar to the task described in Newton and de Villiers (2007), this task utilized a location-change false belief task paradigm. Unlike the Newton and de Villiers (2007) task, which used a single trial (between-subjects design), the current task used multiple trials (within-between-subjects design).

Design. Each participant completed eight trials in total, with two trials per belief/load combination (false/verbal; false/spatial; true/verbal; true/spatial). Participants were randomly

assigned to one of four orders (all participants completed the same trials). Trial presentation order was pseudorandom and counterbalanced by group, with the first trial type balanced across groups to permit comparison of the first trial response only. Responses were measured via button press; performance was recorded for accuracy (correct or incorrect) and response RT (time from stimulus onset to button press).

Primary belief-reasoning task. The belief-reasoning task comprised eight one-minute live-action video trials. In each trial, a young woman ("Annie") hides an apple in Location A, after which a character dressed in a distinctive blue coverall (the Custodian) moves the apple to Location B. In *true belief* trials, Annie watches while her apple is moved; in *false belief* trials, Annie leaves the room and does not see the final location of the apple. Each trial ends with two images: a) Annie searching in Location A (where she originally left her apple) and b) Annie searching Location B (where the Custodian moved the apple). For true belief trials, Location B was correct; for false belief trials, Location A was correct. Participants selected the "correct ending" by pressing a key with their right or left index fingers to select the image on the corresponding side. Participants were instructed to respond as quickly as possible while being as accurate as possible. Photo location was counterbalanced by trial. All videos used the same two actors and target object (apple), but depicted four different rooms and hiding locations. See Figure 2 for images from a sample false belief trial.

Figure 2. Depiction of a sample false belief trial video: Annie places her apple under the coat and leaves the room. The Custodian enters the room and moves the apple from under the coat to the garbage can. Annie then re-enters the room. The final screen is then presented, showing two possible endings to the scene: Annie looking under the coat or in the garbage can. Participants are instructed to select the photo that shows where Annie will look for her apple.

Secondary load task. The belief-reasoning videos were accompanied by a 1-second metronome (a computer recording of a brief "click" sound). Participants performed one of two secondary tasks in time to the metronome: in the *verbal load* condition (four trials), participants recited the days of the week at a rate of one word per second; in the *spatial load* condition (four trials), participants tapped with the left hand in a sequence of four blocks (in a figure eight configuration) at a rate of one tap per second.

Materials. The task was programmed and run using SuperLab 5 software. The experiment was conducted on a 13" MacBook Pro. Color-coded tactile stickers were placed on the "o" and "p" keyboard keys to facilitate accurate responding. The four blocks used for the spatial tapping task were 1.5 cubic inches and were affixed to a foam board in a 4 by 6 inch rectangular configuration.

Procedure. Testing was conducted either at the participant's home or at the University of Connecticut. Participation involved two sessions each lasting three hours. Participants completed the experimental task as well as several standardized measures (detailed in the Measures section). As part of a larger study, participants completed several additional standardized tasks and experimental measures not discussed here.

Prior to beginning the experimental task, participants completed extensive task training. First, participants received instructions on the verbal and spatial load tasks and practiced each of these tasks in time to the metronome until they were comfortable. Participants also practiced selecting images on the screen using the keyboard response buttons. No participant had difficulty understanding or carrying out these tasks.

To become familiar and comfortable with study procedures, participants completed two practice trials: they first watched a video of a causal event (a person's hand spilling a glass of

milk, followed by an image of a full glass and an image of an empty glass) while saying days of the week (verbal load) or tapping the blocks (spatial load), and then selected one of the two images using the response buttons. No participants had difficulty understanding the task or selecting the correct response for the training trials.

Participants were next instructed on the belief-reasoning task. They viewed a series of images while the experimenter recited the following script:

> "This story is about a girl named Annie. This is Annie. [*Cartoon picture of Annie*.] Today, Annie brought her favorite snack to school – a green apple. [*Picture of apple.*] Annie is looking for a safe place to hide her apple until lunch. But there's a problem: whenever Annie hides her apple, the Custodian moves it! [*Cartoon picture of Custodian*.] Sometimes, Annie is in the room when the Custodian moves her apple, but sometimes she's not. You will see two photos on the screen that show two possible endings to the story. Your job is to pick the ending that shows where Annie *thinks* her apple is."

Participants were then tested for task comprehension prior to starting the experiment; testing did not proceed until the participant demonstrated comprehension of the task. All participants were able to achieve task comprehension prior to beginning the task.

Participants then completed the eight experimental trials. At the end of each trial, participants responded by pressing the "o" key with the right index finger or the "p" key with the right middle finger.

Results

Between-subjects independent variables included group (ASD, TD) and task order (1, 2, 3, 4). Within-subjects independent variables included belief (true, false) and load (verbal, spatial). Dependent variables included task *accuracy* and *reaction time* (RT). Accuracy data was calculated as number of correct trials divided by total number of trials multiplied by 100, yielding a percent accuracy score.

Preliminary Analyses. Dependent variables (accuracy and RT) were checked for outliers and missing data. Consistent with prior studies, RTs faster than 100 ms (see Whelan, 2010) and outlier RTs, defined as 2.5 *SD* above or below the mean (see Balota & Spieler, 1999) were removed from the data; 9 trials were removed[1].

Assumptions of normality and homogeneity of variance were examined. Accuracy data met the assumption of homogeneity of variance (Levene statistic $p = .65$), but were non-normally distributed with negative skew and negative kurtosis (Shapiro-Wilk *p*s < .05); these data could not be normalized using standard transformations. As a result, main and interaction effects on accuracy were first examined using a three-way repeated-measures mixed-model ANOVA (group x load x belief). Post-hoc nonparametric tests (Wilcoxon Signed Rank tests; Mann-Whitney U tests) were used to explore any significant main or interaction effects.

RT data were normally distributed (Shapiro-Wilk *p*s > .60) but violated the assumption of homogeneity of variance between groups (Levene statistic $p < .05$); a logarithmic transformation was applied to the data before formal analyses were performed, after which the assumption was met. A three-way repeated-measures mixed-model ANOVA (group x load x belief) tested for main and interaction effects. Post-hoc *t*-tests probed load-by-belief group differences.

Effect sizes were calculated with partial eta squared (η_p^2), which describes the proportion of variance attributable to a given effect after partialling out non-error sources of variance (Cohen, 1988). Although the groups differed in VIQ, because VIQ is inherently related to the independent variable (diagnostic group), it was not included as a covariate (see Dennis et al., 2009).

[1] After outliers were removed, overall RTs were calculated for each participant as the mean of the remaining trials.
[2] Geometric mean (GM) is the back-transformed value of the log-of a set of numbers, using the product of their values.
[3] Based on piloting of the task, these instructions were necessary for participants to establish a baseline understanding of the task.

To examine within-group correlations, Pearson correlations were used (conventions for effect size, *r*: small = .10; medium = .30; large = .50). When assumptions of normality were violated (for RT data), Kendall's τ nonparametric analyses were conducted.

Order effects: Accuracy. Order effects were examined independently and in relation to the other variables of interest: group, load, and belief. There were no main effects of order, $F(3,24) = .35$, $p = .79$, $\eta_p^2 = .04$., and no two-way interactions (order x group, order x load, or order x belief), all *p*s > .27. However, there was a significant three-way interaction of order x group x load, $F(3,24) = 7.03$, $p < .01$, $\eta_p^2 = .47$ (Figure 3). Post-hoc analyses revealed three-way interactions between orders 1 and 2, $F(1,12) = 14.6$, $p < .01$, $\eta_p^2 = .53$. This three-way interaction was driven by a two-way load x group interaction, showing that, within order 1 (in which spatial trials were 2, 4, 6, and 8, and verbal trials were 1, 3, 5, and 7), the difference between spatial load and verbal load among TD participants (among whom accuracy was numerically but not statistically greater under spatial load, $p = .18$) was significantly different than the difference between spatial load and verbal load among ASD participants (among whom accuracy was numerically but not statistically greater under verbal load, $p = .21$), $F(1,12) = 9.80$, $p < .05$, $\eta_p^2 = .62$. Between order 1 and 2 (in which spatial trials were 2, 4, 6, and 8, and verbal trials were 1, 3, 5, and 7), no other two-way interactions or main effects were significant, all *p*s > .16. The post-hoc analyses also revealed a significant three-way interaction of order x group x load between orders 1 and 4 (in which spatial trials were 2, 4, 7, and 8, and verbal trials were 1, 3, 5, and 6), $F(1,12) = 13.4$, $p < .01$, $\eta_p^2 = .55$. Between orders 1 and 4, there was also a two-way group x order interaction within verbal trials, $F(1,12) = 8.78$, $p < .05$, $\eta_p^2 = .44$, with TD participants in order 4 being more accurate than ASD participants in that order, $F(1,12) = 9.55$, $p < .05$, $\eta_p^2 = .66$. There were no four-way interaction effects, all *p*s > .05. Because the order effects in orders

1, 2, and 4 were nonsystematic (e.g. there were no clearly aberrant accuracy profiles; accuracy in order 3 did not differ from the other order conditions), all *ps* > .20, we believe these results reflect idiosyncratic accuracy data. Moreover, dividing the sample by order would yield only four participants per cell, reducing power and increasing the risk of Type II errors. All data were therefore collapsed across order in subsequent analyses.

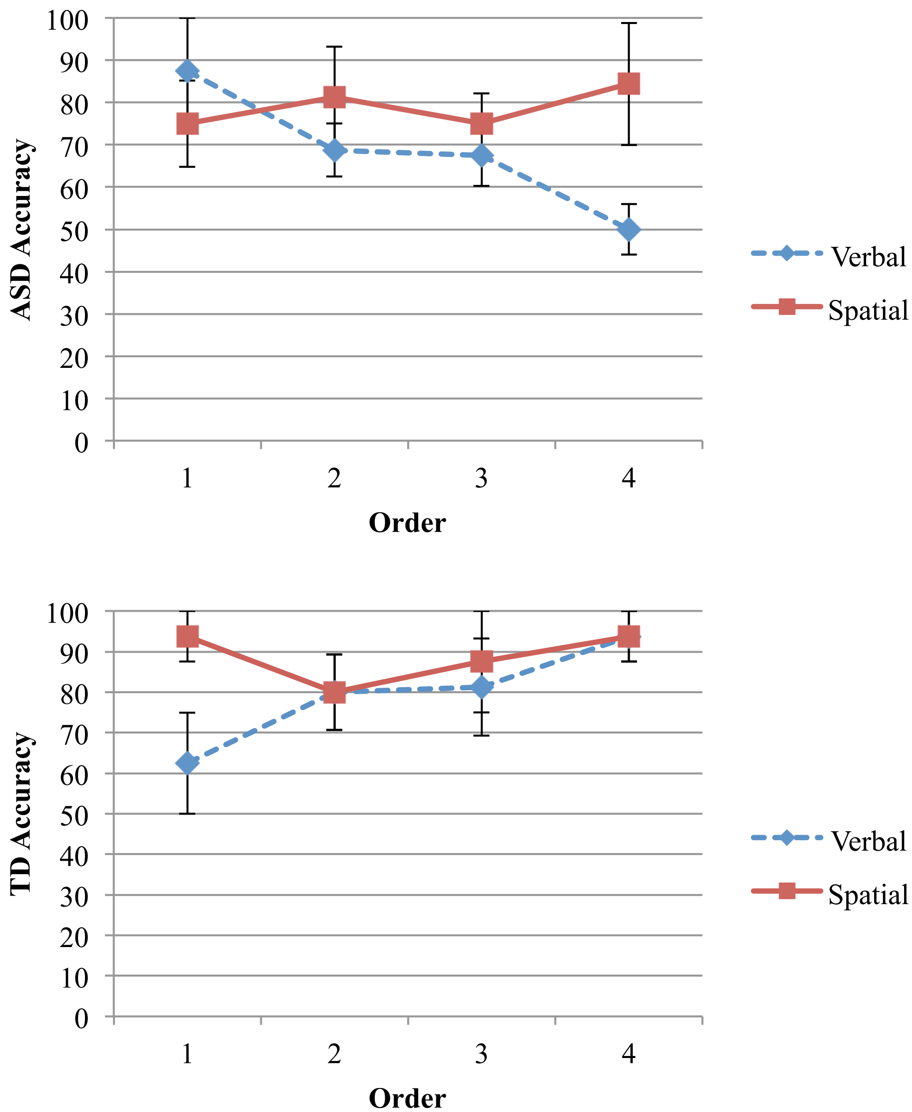

Figure 3. Accuracy data for order effects. Top figure displays ASD data; bottom figure displays TD data. Error bars represent standard errors of the means.

Order effects: RT. There was no main effect of order, $F(3,28) = .18$, $p = .08$, $\eta_p^2 = .25$. There was a significant interaction of order x belief, $F(3,28) = 3.76$, $p < .05$, $\eta_p^2 = .32$ (Figure 4). Post-hoc analyses revealed an order x belief interaction for orders 1 and 3, $F(1,14) = 5.43$, $p < .05$, $\eta_p^2 = .28$: in order 1 (in which true belief trials were 1, 4, 7, and 8, and false belief trials were 2, 3, 5, and 6), participants were faster on false belief trials as compared to true belief trials, $F(1,7) = 8.58$, $p < .05$, $\eta_p^2 = .55$. The difference for false versus true belief trials was not significant in order 3, $F(1,7) = 1.10$, $p = .33$, $\eta_p^2 = .14$. There were no additional two-way, three-way, or four-way interactions, all *p*s > .37. Because these order effects did not involve diagnostic group differences, and because orders 2 and 4 did not differ from orders 1 and 3, these results reflect nonsystematic findings. All orders were therefore collapsed in subsequent analyses.

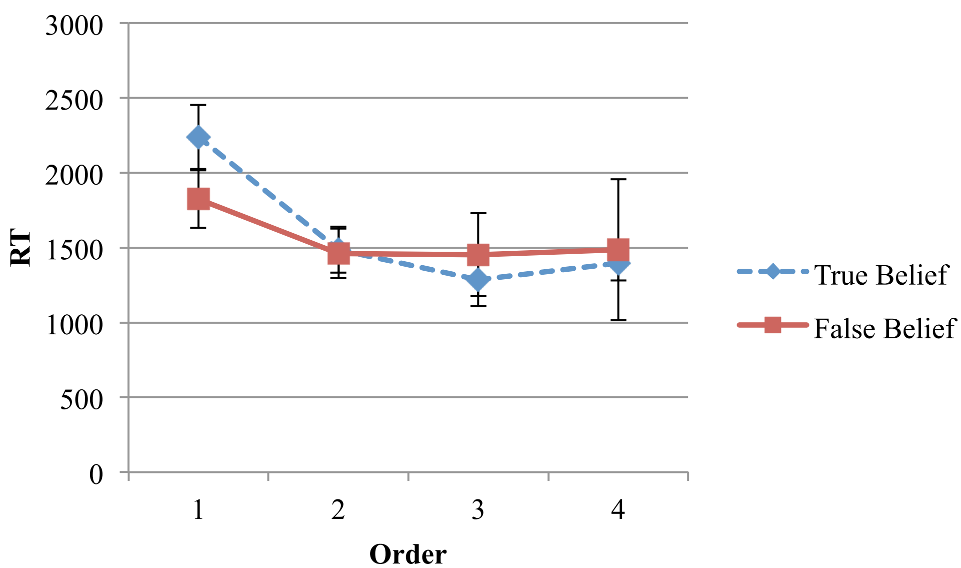

Figure 4. RT data for order effects. Error bars represent standard errors of the means.

Accuracy. We hypothesized that verbal load would interfere more with false belief task performance for the TD group than for the ASD group. As such, we examined effects of group

(TD; ASD), load (verbal; spatial), and belief (false; true) using a three-way mixed-model ANOVA (group x load x belief).

Overall, task accuracy was high: mean accuracy was 81%, with a range of 50% to 100%; chance performance was 50%. Analyses revealed no main effect of group, $F(1,30) = .96$, $p = .34$, $\eta_p^2 = .03$: overall accuracy rates were comparable for participants with ASD (*M* = 77.5%; *SD* = 18.4) and TD (*M* = 83.8%; *SD* = 18.1). Collapsing across groups, Wilcoxon Signed Rank tests revealed no main effect of belief type, z = -1.07, p = .28. There was, however, a main effect of load, such that across groups, participants were significantly less accurate under verbal load (*Md* = 75%; *SD* = 22) than under spatial load (*Md* = 100%; *SD* = 19), $z = -1.98$, $p < .05$, with a small to medium effect ($r = .25$)

There was no interaction between group and belief, $F(1,30) = .17$, $p = .69$, $\eta_p^2 = .01$. Importantly, accuracy on false belief trials specifically was equivalent between the ASD group (*M* = 80%; *SD* = 27) and TD group (*M* = 90%; *SD* = 24). There was no interaction between group and load, $F(1,30) = .31$, $p = .58$, $\eta_p^2 = .01$, and there was no was group x load x belief interaction, $F(1,30) = 1.09$, $p = .31$, $\eta_p^2 = .04$.

The main effect of load was explored further by examining the load x belief interaction: this interaction was not significant, $F(1,30) = .34$, $p = .56$, $\eta_p^2 = .01$. These findings suggest that, though participants in both groups were less accurate under verbal load, this verbal interference effect was not specific to false-belief performance. Rather, participants experienced verbal interference across both true and false belief trials. These results suggest that task accuracy was generally high, though not at ceiling, with no group differences, and that verbal load had a greater impact on task performance than spatial load for both diagnostic groups, regardless of trial type. Accuracy data are presented in Table 2 and Figure 5.

Table 2. *Percent accuracy by diagnostic group, load, and belief*

	TD		ASD	
	M	SD; Range	M	SD; Range
Verbal Load				
False Belief	88	(28) 0-100	77	(32) 0-100
True Belief	70	(43) 0-100	73	(37) 0-100
Spatial Load				
False Belief	91	(19) 50-100	83	(30) 0-100
True Belief	85	(34) 0-100	77	(32) 0-100

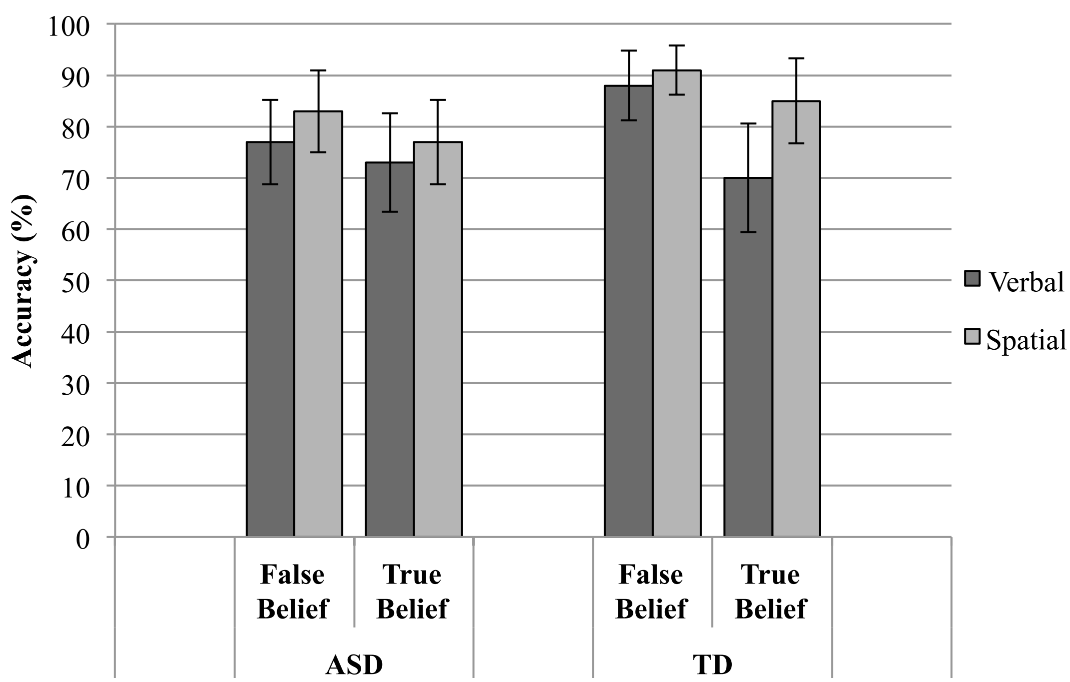

Figure 5. Mean percent accuracy as a function of belief condition and load condition for ASD and TD participants. Error bars represent standard errors of the means.

Reaction time. Because implicit, on-line responses may be more sensitive to individual differences, we examined RT as a second metric of task performance. We conducted a second repeated-measures mixed-model ANOVA (group x load x belief), using log-transformed RT as the dependent variable.

The overall geometric mean (GM)[2] RT was 1567 ms, with a range of 489 to 5370 ms. There was a significant main effect of group on RT, $F(1,30) = 10.45$, $p < .005$, $\eta_p^2 = .26$, such that ASD participants (*GM* = 1922 ms; 95% CI [1547, 2388]) were significantly slower than TD participants (*GM* = 1309 ms; 95% CI [1147, 1493]). In order to examine factors driving this group difference, we tested whether group interacted with load and belief conditions. None of these interactions were significant, all *p*s > .10. These findings suggest that ASD participants responded more slowly in general, and that this slower RT was not driven by interference load or trial type.

There was no main effect of load, $F(1,30) = .34$, $p = .56$, $\eta_p^2 = .01$, or belief, $F(1,30) = .34$, $p = .56$, $\eta_p^2 = .01$, and no load x belief interaction, $F(1,30) = .34$, $p = .56$, $\eta_p^2 = .01$, suggesting that neither load, nor trial type, nor their interaction affected RT. RT data are presented in Table 3 and Figure 6.

[2] Geometric mean (GM) is the back-transformed value of the log-of a set of numbers, using the product of their values.

Table 3. *Reaction time (ms) by diagnostic group, load, and belief*

	TD		ASD	
	GM	95% CI; Range	GM	95% CI; Range
Verbal Load				
False Belief	1251	[1056, 1483]	1930	[1417, 2629]
		489-2089		603-5370
True Belief	1363	[1162, 1599]	1638	[1310, 2047]
		871-2455		832-2752
Spatial Load				
False Belief	1241	[1085, 1419]	1902	[1577, 2294]
		832-2042		933-3477
True Belief	1310	[1121, 1532]	2029	[1526, 2635]
		661-2399		871-4467

Table reports geometric means, followed by 95% CIs, followed by ranges

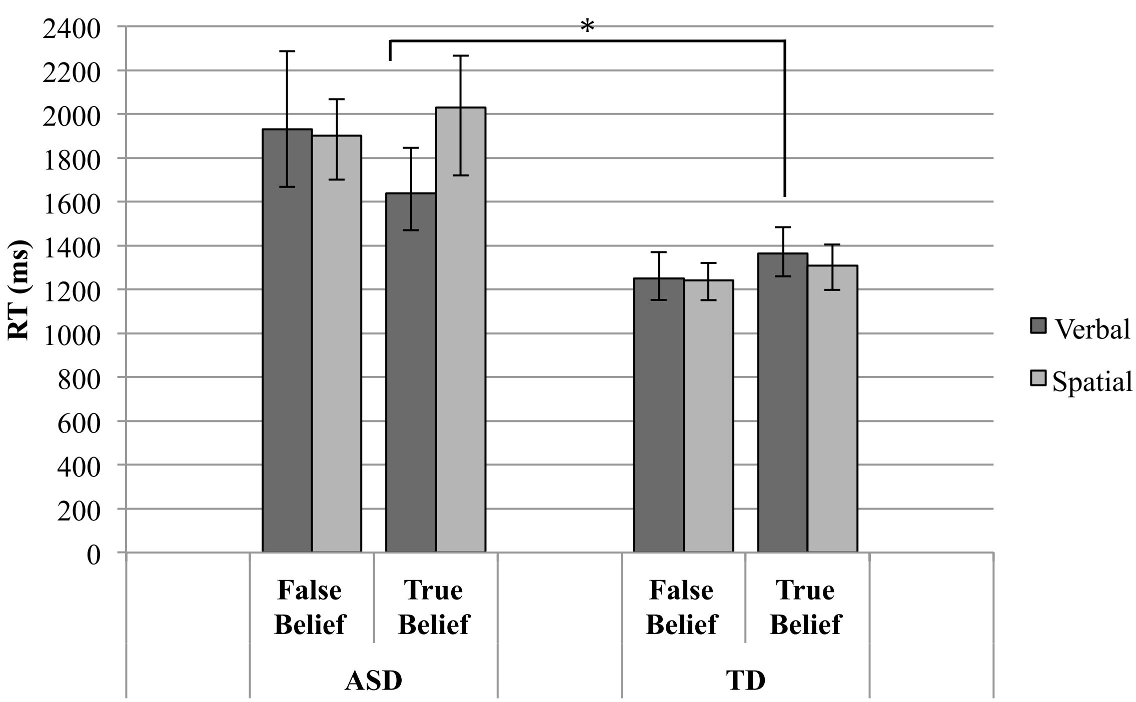

Figure 6. Geometric mean RT as a function of belief condition and load condition for ASD and TD participants. Error bars represent standard error of the geometric means. Asterisks indicate significant differences between conditions, $p < .005$.

Trial 1 accuracy and RT. Between-subjects ANOVAs were used to examine trial 1 accuracy and RT for trial condition (belief/load) by group. There were no main or interaction effects of trial condition by group on trial 1 accuracy or RT, all *p*s > .10. It is important to note that, given the between-subjects analyses necessary for examining trial 1 differences, however, each cell contained only 4 participants. Results should therefore be interpreted with caution due to the increased risk of Type II errors.

Correlational analyses: language ability. Because verbal load did not specifically interfere with false belief, we looked for a direct relationship between language and belief reasoning by examining language ability correlates of performance on false belief versus true belief trials. Within each diagnostic group, we examined correlations between CELF Core Language, VIQ, and NVIQ scores and false belief RT, as well as true belief RT. Due to violations of assumptions of normality, as well as high frequency of tied ranks within the data, we used nonparametric analyses. Data are presented in Table 4 and Figure 7.

There was no significant correlation between CELF Core Language and false or true belief RT within either group, all *p*s > .20. Within the ASD group, however, VIQ correlated with RT for false belief trials, $r_\tau(13) = -.40$, $p < .05$, but not true belief trials, $r_\tau(13) = -.14$, $p = .48$, such that, among ASD participants, higher VIQ was associated with faster RT for false belief trials. Within the TD group, there was no relationship between false belief trials, $r_\tau(15) = .01$, or true belief trials, $r_\tau(15) = -.16$, *p*s > .41. In order to examine whether this relationship between RT and VIQ within the ASD group was specific to verbal ability, we also examined the relationship between NVIQ and false belief RT; there was no correlation within the ASD group, $r_\tau(13) = .08$, $p = .69$, or the TD group, $r_\tau(15) = .02$, *p*s > .69. These findings suggest that, within the ASD group, VIQ predicted faster processing of false belief trials above and beyond more

general task demands. Within the TD group, however, VIQ was not associated with speed of false belief processing. Data are presented in Table 4.

Table 4. *Correlation table of CELF, VIQ, and NVIQ versus false belief RT, and true belief RT within ASD and TD groups. Data are reported as Kendall's τ* values.

	ASD	
	False Belief RT	True Belief RT
CELF Core Lang.	-.22	-.05
VIQ	-.40*	-.14
NVIQ	.08	-.04
	TD	
	False Belief RT	True Belief RT
CELF Core Lang.	.00	-.08
VIQ	.01	-.16
NVIQ	.02	.16

* $p < .05$

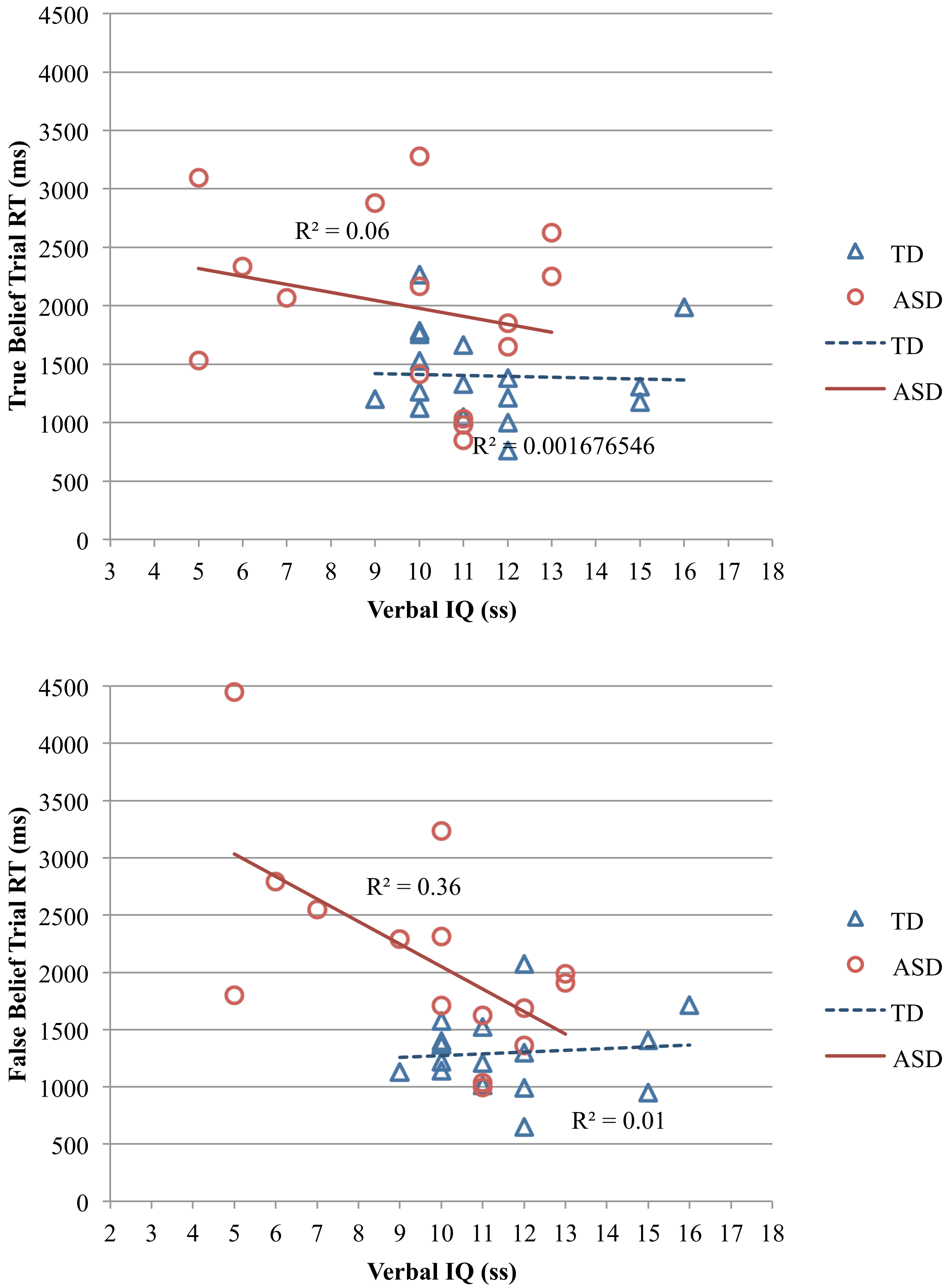

Figure 7. Scatterplots and regression lines of VIQ by True Belief RT (top figure) and by False Belief RT (bottom figure) by diagnostic group.

Correlational analyses: executive functions. Given evidence suggesting that executive functions, particularly inhibitory control, are associated with false belief task performance, we examined the relationship between general EF as well as inhibitory control and belief reasoning within each group. Within each group, we examined correlations between BRIEF General Executive Composite (for which lower scores indicate stronger EF) and D-KEFS Color-Word Interference Inhibition performance (for which higher scores indicate stronger EF) versus false belief and true belief RT. Due to violations of assumptions of normality and high frequency of tied ranks within the data we again used nonparametric analyses. There were no significant correlations for either the BRIEF GEC or D-KEFS Color-Word Inhibition with false or true belief RT within either group, all *p*s > .05. Data are presented in Table 5.

Table 5. *Correlation table of D-KEFS Color-Word Interference and BRIEF versus false belief RT, and true belief RT within ASD and TD groups. Data are reported as Kendall's τ* values.

	ASD	
	False Belief RT	True Belief RT
D-KEFS C-W Inhib.	.03	.09
BRIEF GEC	-.01	.03
	TD	
	False Belief RT	True Belief RT
D-KEFS C-W Inhib.	.21	.38
BRIEF GEC	.19	.22

* $p < .05$

Study 1 Summary

We used an experimental dual task paradigm in which participants completed false and true belief tasks while engaging in a verbal load task (to suppress inner speech), or a spatial load task (designed to preserve inner speech). We predicted that verbal load would interfere with false

belief, and that language abilities would be associated with more efficient false belief processing, in the TD group but not the ASD group. There was no general false belief deficit in this highly verbal group of youth with ASD for false belief task performance. Contrary to our hypotheses, we found no verbal load interference of the false belief task in either the ASD or the TD group. We also predicted that EF would be associated with false belief processing within both groups. We found, however, no evidence to support this relationship within either group. We did find that, regardless of diagnostic group, participants were generally less accurate on the belief-reasoning task during the verbal load task as compared to spatial load task. To rule out the possibility that this performance difference reflects greater difficulty of the verbal load task as compared to the spatial load task, based on an approach used in Hegarty, Shah, and Miyake (2000), we compared load task performance using a secondary (dual-task) methodology.

Study 2: Perceptual Matching Dual Task

To compare the difficulty of the verbal versus spatial interference tasks, we conducted a second study utilizing a Perceptual Matching task, which involved low-level visual scanning and has been shown to minimally involve verbal mediation and central executive functions (Hegarty et al., 2000).

Methods

Participants. Participants were 32 undergraduates at the University of Connecticut who participated for course credit. All participants were native English speakers and had no significant visual or auditory impairments that would impede study participation. This study was approved by the University of Connecticut Institutional Review Board. Participants provided written consent for study participation.

Perceptual matching dual task. The primary task was a "perceptual matching task," adapted from a task used in Hegarty et al. (2000), which required primarily visual scanning to match line drawings. Given that the aim of this secondary study was to test the difficulty of the two dual tasks used in Study 1, this primary task was selected to draw on low-level cognitive functions (e.g., attention) but place minimal demands on executive functions and require minimal verbal mediation. The task was adapted from the Identical Pictures Test in the Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman, & Dermen, 1976).

In this task, participants were instructed to match a target figure (on the left side of the horizontal line; Figure 7) to one of four figures on the right as quickly and as accurately as possible, by pressing one of four buttons corresponding with the four fingers of the right hand. There were 25 randomized trials in each block; all trials were comprised of different line drawings. Responses were recorded for accuracy and RT. Responses were measured by button press; performance was recorded for accuracy (correct or incorrect) and reaction time (RT; as measured as time from onset of stimulus presentation to button press).

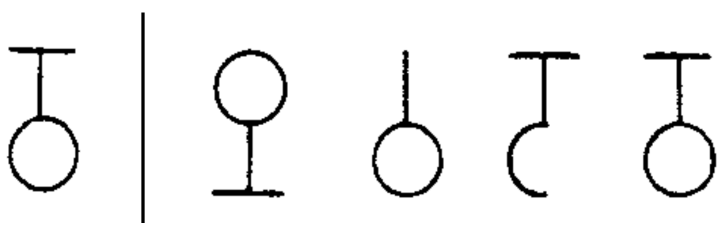

Figure 7. Example trial from the Perceptual Matching Task.

Secondary load task. The secondary task was the same as in Experiment 1: while completing the perceptual matching (primary) task, participants performed a dual task (imposing a verbal load by reciting days of the week, or a spatial load by tapping a sequence of four blocks

with the left hand. Unlike in Experiment 1, participants in this experiment performed each dual task continually throughout the 25 trials of each block.

Design. Participants were randomly assigned to one of two counterbalanced orders. In both orders, participants completed two blocks of 25 primary task trials for a total of 50 trials: the 25 trials within each block were randomized. In order 1, participants first completed the verbal load trials, followed by the spatial load trials; this was reversed in order 2.

Materials. The task was programmed using SuperLab 5. The experiment was conducted on a 21.5-inch iMac. Responses were given using a Cedrus RB-740 response pad. The four blocks used for the spatial load (tapping) task were affixed to a foam board in a 4 by 6 inch rectangular configuration.

Procedure. The experiment was conducted in a quiet lab room at the University of Connecticut and lasted 20-25 minutes. Before beginning the task, participants first completed a brief training. The secondary task training was identical to that presented in the belief reasoning experiment: participants received instructions for the verbal and spatial load tasks and practiced them in time to a metronome until they were comfortable. They also practiced pressing the four buttons with the right hand that corresponded to the four possible responses. Before starting the experiment, the participants completed several practice trials of the perceptual matching task, first without a secondary load task and then with each load task. They then completed the two blocks of 25 experimental trials.

Results

Dependent variables were task *accuracy* and *reaction time* (RT). Accuracy data was calculated as number of correct trials divided by total number of trials multiplied by 100, yielding a percent accuracy score.

Preliminary Analyses. Dependent variables (accuracy and RT) were checked for outliers and missing data. As in Study 1, outlier RTs, defined as 2.5 *SD* above or below the mean, and RTs faster than 100 ms were removed from the data. Assumptions of normality and homogeneity of variance were examined.

Accuracy data were non-normally distributed with negative skew (Shapiro Wilk $p < .05$); a reflected logarithmic transformation was applied before formal analyses were performed. RT data were non-normally distributed with positive skew and positive kurtosis (Shapiro-Wilk *p*s < .05); a logarithmic transformation was applied. After transformation, both measures met statistical assumptions. Two-way repeated-measures mixed-model ANOVAs (load x order) tested for main and interaction effects of both Accuracy and RT data. Post-hoc *t*-tests were used to examine load-by-order interactions. Effect size was calculated with partial eta squared (η_p^2).

Accuracy. We examined effects of load (verbal, spatial) and order (1, 2) using a two-way mixed-model ANOVA, using reflected-log-transformed percent accuracy as our dependent variable, with effect of load condition on accuracy as our primary contrast of interest.

The overall (geometric) mean task accuracy was 90%, with a range of 68% to 100%; chance performance was 25%. Analyses revealed no main effect of load, $F(1,30) = 2.59$, $p = .19$, $\eta_p^2 = .08$, or order, $F(1,30) = 1.84$, $p = .19$, $\eta_p^2 = .06$, and no load and order interaction, $F(1,30) = .78$, $p = .38$, $\eta_p^2 = .03$, suggesting that neither the load task nor the task order affected accuracy.

Reaction time. We next examined effects of load and order on log-transformed RT, again using a two-way mixed-model ANOVA. The effect of load condition on RT was the primary contrast of interest.

The overall (geometric) mean RT was 1824 ms. Analyses revealed a significant main effect of load, $F(1,30) = 6.48$, $p < .05$, $\eta_p^2 = .18$, where participants had significantly slower RTs

under spatial load (*GM* = 1880; 95% CI [1734, 2038]) than under verbal load (*GM* = 1759; 95% CI [1635, 1892]). There was also a load x order interaction, $F(1,30) = 10.92$, $p < .005$, $\eta_p^2 = .27$; a post-hoc paired-samples *t*-tests revealed that only participants in order 2 (who completed spatial load trials first) were slower under spatial load as compared to verbal load, $t(31) = -4.23$, $p < .001$, Cohen's $d = .72$. There was no difference in RT for verbal load trials, $t(30) = .21$, p = .84, Cohen's $d = .08$. Given that participants in order 2 completed the spatial block first and the verbal block second, and that participants who completed the spatial block first were slower on that block, these findings suggest a benefit from practice effects for spatial load trials only.

Study 2 Summary

Results indicate that accuracy in a low-level, non-verbally-mediated task did not differ for verbal and spatial load tasks, suggesting that these interference tasks did not *differentially* interfere with accuracy in Study 1. Results did reveal, however, that RT on the Perceptual Matching task did differ for verbal versus spatial interference, such that participants were slower to respond during the spatial interference task, particularly when they encountered that task first. This suggests an effect of practice. These data also indicate that slower RT on the Study 1 Belief-Reasoning task under verbal load was not due to it being more demanding. In fact, this study suggests that the verbal interference task was *less* taxing of general cognitive and attention functions than the spatial load task.

Discussion

The current study examined the relationship between language and ToM in adolescents with ASD and TD. To investigate whether or not verbal load interfered with belief reasoning, participants completed an experimental dual task paradigm in which they completed false belief tasks while engaging in a verbal load task, designed to suppress inner speech, and a spatial load

task. Given evidence that ToM is verbally mediated in TD, and that people with ASD rely less on verbal mediation in general, we predicted that verbal load would interfere with false belief in the TD group but not the ASD group. We also examined predictors of false belief by testing associations between false belief processing (as measured by RT on false belief trials) and language ability (as measured by CELF Core Language and VIQ as compared to NIVQ) within each group, as well as inhibitory control (as measured by D-KEFS Color-Word Interference), and general executive functions (as measured by the BRIEF Global Executive Composite). Lastly, to compare the cognitive load imposed by a verbal versus a spatial dual task, a second study examined performance in a non verbally-mediated (picture matching) task under verbal versus spatial loads. This allowed us to test whether either load was more cognitively demanding, or whether the effects of verbal load were specific to the belief-reasoning task.

Our findings revealed comparable task accuracy between groups, for both false belief and true belief trials, and no significant interactions within or between groups for load task or belief status. Findings indicated a main effect of group on RT, such that ASD participants were slower overall than their TD counterparts. There were no significant between-group interaction effects, and no effects of load task or belief status. Thus, contrary to our hypothesis, there was no performance decrement (in terms of accuracy or RT) on false belief trials under verbal load for either group. Regarding correlational analyses, there were no associations between false belief RT and CELF Core Language scores within either group. There was an association between false belief RT and VIQ for the ASD but not TD group; in contrast, false belief and NVIQ were uncorrelated. There were no correlations between EF measures and false belief RT within either group.

Below, we review these findings in light of broader issues regarding ToM as a construct,

and discuss criticisms regarding the false belief task as a metric of ToM. We also explore the finding of an association between language false belief performance within the ASD group in the context of a two-systems account of ToM.

False belief performance in ASD. In the current study, TD and ASD participants demonstrated comparably strong performance on false belief tasks, with accuracy rates of 88% and 80% respectively. Thus, ASD participants in our study showed no generalized deficit in false belief task performance. This finding contradicts literature suggesting broad deficits in ToM in people ASD. Importantly, however, the verbal ages among our ASD participants in our study were well beyond the verbal age associated with successful false belief task performance (Wellman, 1990). The current results provide compelling evidence that, contrary to common beliefs, ASD individuals with intact language skills have globally intact false belief reasoning. As discussed below, however, it is important to note that false belief tasks represent only one index of ToM.

Verbal interference of false belief. Overall, results from the belief-reasoning dual task were inconsistent with our hypothesis. Contrary to predictions, we found no evidence that verbal load interfered with false belief *per se* in either the ASD group or the TD group. Participants under verbal load were no less accurate for false belief than for true belief trials, nor were there RT differences. In fact, accuracy rates on false belief trials under verbal load were 77% and 88% for ASD and TD groups, respectively. These findings are inconsistent with Newton and de Villiers (2007), in which verbal loads disrupted false belief task performance (42% correct) while sparing true belief performance (91% correct).

Given their findings, Newton and de Villiers (2007) concluded that inner speech is required for false belief task performance. If so, how might we explain the success of

participants in the current task, who presumably had little access to inner speech under verbal load? There are several possibilities.

First, the experimental design of the current study differed from that of Newton and de Villiers (2007) in several crucial ways. While Newton and de Villiers (2007) used a between-subjects design (i.e., every participant completed only one trial), we used a within-subjects design (i.e., every participant completed two true belief and two false belief trials under both load conditions); this was a necessary design feature in a study including a clinical population, where sample sizes are necessarily smaller. The absence in the current study of false belief-specific verbal interference may reflect, in part, the fact that participants completed multiple belief trials in succession. Although each trial was set in a different room with a different hiding place for Annie's apple, there was only one substantive manipulation: whether or not Annie was in the room when the Custodian moved the apple from one location to another. Based on subjective observations as well as participant debriefing after the task, we suspect that most participants were able to recognize this manipulation and "get into set," such that they were able to circumvent the need to verbally mediate conflicting belief representations repeatedly for each trial. For example, during trials in which Annie left the room (which was the signature manipulation of false belief trials), many participants (both ASD and TD) reported that they simply "held in mind" the original apple location and "tuned out" the location change, knowing that the original location would always be the correct answer when Annie left the room. This heuristic allowed participants to bypass verbally mediated false belief reasoning, thus minimizing the impact of verbal interference on task performance.

Moreover, in Newton and de Villiers (2007), it is unclear what, if any, task instructions were provided to participants prior to beginning the task. The current experiment, on the other

hand, provided verbal and written task explanations to participants, including specific mention of Annie either being in the room or not, and that the goal was to pick the ending "where Annie thinks her apple is."[3] It is likely that these explicit instructions effectively helped participants anticipate and verbally reason about belief representations prior to beginning the task at all, potentially minimizing their reliance on inner speech.

The verbal and spatial interference stimuli also differed between studies. In the current experiment, participants tapped a sequence of blocks or said a sequence of words, in time to a metronome. In Newton and de Villiers (2007), however, participants were required to either perform brief "call and response" tapping sequences or verbally echo a complex verbal narrative in real time.[4] It is possible that the attentional and cognitive demands of the interference tasks, particularly the verbal interference tasks, differed greatly between studies. Finally, Newton and de Villiers's (2007) task included two location changes (e.g., location A to B to C), whereas the current task only included one location change (e.g., location A to B). The need for participants to track the object location among three locations in Newton and de Villiers (2007) likely increased the executive load (e.g., attention and working memory) of their task. Together, these task differences may help to account for the discrepant findings of the current study and those of Newton and de Villiers (2007).

These discrepant findings highlight inconsistencies in the meaning and use of the concept of *representational* false belief. All participants in both studies likely possessed the *capacity* to represent false belief.[5] In Newton and de Villiers (2007), while participants' capacity for false

[3] Based on piloting of the task, these instructions were necessary for participants to establish a baseline understanding of the task.
[4] Pilot testing with the dual task stimuli from Newton and de Villiers (2007) revealed that our participants were unable to attend to the videos during verbal shadowing. They also reported that the verbal shadowing task was significantly more difficult than the rhythmic shadowing task.
[5] In our task, this was confirmed when testing for task comprehension and in debriefing responses.

belief was presumably intact, the verbal interference task, by hindering inner speech, prevented participants from *applying* false belief representations. The findings of Newton and de Villiers (2007) may have less to do with false belief *per se*, and more to do with online processing of tasks involving multiple representations. The experimental designs of these belief-reasoning studies, therefore, fundamentally affected their respective results. It is concerning that relatively subtle differences in experimental methodology and task design between these two studies led to fundamentally different conclusions about false belief and ToM. As we will later discuss, this point parallels concerns raised by several authors about the validity of using false belief tasks as a proxy for ToM.

General verbal interference. Although verbal load did not interfere with false belief performance *per se*, Study 1 revealed that participants were less accurate on all trials — both false and true belief — when under verbal load as compared to spatial load, regardless of diagnostic group. This does not reflect a simple difference in load task difficulty; Study 2 revealed that the two load tasks exerted similar constraints on a simple picture-matching task, with RT differences suggesting that the spatial load task was the more difficult one. It is possible that the motor demands shared between the spatial load task (performed with the left hand) and the mode of responding (pressing with one of the four digits on the right hand) contributed to the slower performance observed under spatial load. Nonetheless, it appears that, in terms of general attentional demands, the verbal load task used in these studies was not more taxing than the spatial load task. However, In Study 1, the false *and* true belief-reasoning tasks were both more affected by verbal load, suggesting that participants relied upon verbal mediation to solve both false *and* true belief tasks in this study.

The current experiment included four true belief trials, which were intended to act as a

"control" condition to which performance on false belief trials could be compared. After observing participants complete the task, however, we believe that the true belief trials likely involved more ToM – and therefore more inner speech – than anticipated, thus making them ineffective controls. Recall that both the false and true belief tasks involved Annie placing her apple in a given location, followed by the Custodian entering the room and moving her apple to a new location, followed by Annie looking for her apple. The only meaningful difference between the true and false belief trials was as follows: in the false belief condition, Annie left the room before the Custodian moved her apple, whereas in the true belief condition, Annie remained in the room. Though it is the case that only the false belief condition involved false belief *per se*, the true belief condition also required participants to track Annie's location, and therefore her belief about the apple's location. In fact, because the true belief trials were intermingled with false belief trials, participants may have been primed to attend to Annie's mental state during the true belief trials. In other words, because participants attended to the true belief task in the context of the false belief task, they were required to track Annie's mental state during both tasks. At the very least, both the true and false belief conditions of the task required similar processing demands: remembering the goal of the task, tracking the apple's location, and inhibiting one response in favor of another. In general, these findings highlight the complexity of the relationship between false belief tasks and ToM.

What is the relationship between false belief and theory of mind? The discrepancy between the current findings and those of Newton and de Villiers (2007) underscores broader methodological and theoretical issues regarding ToM. The false belief task paradigm was originally crafted to index the emergence of ToM among preschool-aged children. Successful false belief task performance requires a person to appreciate that others' actions are determined

not by the true state of the world, but by their mental representations of the world, which may or may not be accurate. An *appreciation* of others' mental representations, however, does not guarantee successful false belief task performance. Indeed, most standard false belief tasks require the explicit *application* of this knowledge, which, as many authors have noted, require cognitive skills above and beyond ToM representation. Bloom and German (2000) argue, for example, that selecting the correct response on a false belief task requires an individual to override useful heuristics about the world. Thus, successful false belief task performance appears to be related to inhibitory control (Leslie & Polizzi, 1998). Participants' ability to inhibit a "default" (true) response in favor of the correct (false) response is inherent to successful performance (Leslie et al., 2004). Bloom and German (2000) also argue that a person's ToM need not entail the ability to reason about false beliefs. As we will discuss in sections below, empirical and everyday evidence points to examples of children younger than three, who would fail standard false belief task, demonstrating appreciation of the minds of others.

There is no doubt that the false belief task is a useful tool. It is important, however, that researchers view the false belief task as a tool for indexing a behavior *associated with* ToM rather than ToM itself. As Bloom and German (2000) state, the false belief task "is a task that taps one aspect of people's understanding of the minds of others; nothing more, nothing less" (p.30).

This discussion also raises a broader question: what is ToM? Is there currently a unifying definition of the construct of ToM? Would such a construct even be meaningful? Apperly (2012) argues that, despite the existence of a consensus regarding the meaning of ToM, there are, in fact, at least three views of ToM implicit in research: a conceptual perspective, which assumes that false belief tasks index a person's conceptual knowledge of ToM; a cognitive perspective,

which assumes that false belief tasks index a person's capacity to contextually apply ToM; and the social individual differences perspective, which assumes that false belief task errors can occur at either the conceptual level or cognitive level of processing, depending upon the individual's specific deficits (e.g., perspective-taking versus inhibitory control). Apperly (2012) ultimately argues that the construct of ToM itself is a straw man, as the notion of such a construct perpetuates the flawed assumption that ToM is a definable cognitive entity rather than an imperfect heuristic for representing the complexities and nuances of the social mind. Given the ill-defined nature of ToM, it is no surprise that identifying – much less quantifying – ToM remains a nebulous task. This issue is reflected in several contemporary studies, which paint a more nuanced picture of ToM development than previously assumed. As we will discuss below, several authors have worked towards a more sophisticated understanding of how people reason about the minds of others.

Two systems of theory of mind. Although children younger than age four consistently fail traditional false belief tasks, more recent evidence, both observational and experimental, suggests that they *do* have the capacity to represent the minds of others.

Young children are capable of attributing goals to other agents (Gergely, Nádasdy, Csibra, & Biro, 1995), engaging in pretend play and understanding pretense (Leslie, 1994), and imitating the intended (but not accidental) actions of others (Meltzoff, 1995). Indirect evidence also suggests that children younger than four have some form of ToM. For example, O'Neill (1996) found that two-year-old children were more likely to reference a toy and point to its location when their parent was aware of the presence of the toy (i.e., witnessed the toy being placed on the shelf) than when the parent was unaware. These two year olds altered their behavior according to the knowledge states of others. In another study, Onishi and Baillargeon

(2005) utilized a violation of expectations false belief paradigm to examine ToM in 15-month-olds. They found that the infants consistently expected another person to reach to a location in accordance with his or her belief about a hidden toy's location rather than the actual location of the toy. A similar study found that, at 13-months, infants looked significantly longer on trials in which an agent searched for an object in a location which was correct but inconsistent with the agent's belief (Surian, Caldi, & Sperber, 2007).

These studies demonstrate that infants as young as 13 months account for others' beliefs when predicting behavior, suggesting that infants possess some sort of representational ToM almost two years before they pass traditional false belief tasks. How do we reconcile these findings with the robust literature showing that children do not pass traditional false belief tasks until age four? Apperly and Butterfill (2009) propose a model of two parallel systems of ToM: an "implicit" system, which is rapid and efficient yet inflexible, and an "explicit" system, which is flexible yet slow, effortful, and mediated by language. Several scholars have, in fact, proposed a dual process theory of cognition more generally. James wrote about "associative" or implicit thinking versus "true" or explicit thinking (James, 1890), and Kahneman (2003) has described the interplay of "intuition," which is unconscious, fast, and automatic, and "reasoning," which is conscious, slow, and governed by language.

Apperly and Butterfill (2009) suggest that the implicit ToM system is pre-linguistic and is present in the first two years of life. The explicit ToM system, on the other hand, recruits domain-general executive functions, such as planning and inhibitory control (Bialystok & Viswanathan, 2009; Leslie et al., 2004), and therefore necessitates language. Tager-Flusberg (2001) also proposed a similar two-system model of ToM, with a social-perceptive (i.e., implicit) ToM system and a social-cognitive (i.e. explicit) ToM system. In this model, the social-

perceptive system enables real-time, automatic social exchanges and intuitive judgments about the minds of others, whereas the social-cognitive system enables effortful, deliberate social exchanges and explicit reasoning about others' mental states, as in the case of complex and executively-demanding false belief tasks. The social-cognitive system requires deliberate and effortful engagement, whereas the social-perceptive system is activated automatically. Typically developing individuals engage both ToM systems, depending upon the context, and often simultaneously. Some scholars have argued that the social-perceptive ToM system is innate (or is acquired in the first few months of life), whereas the social-cognitive system is contingent upon the development of language.

Does language mediate theory of mind in ASD? In the current study, verbal load interfered with task accuracy in both the TD and the ASD group. However, there was an association between VIQ and false belief reaction time within the ASD group, such that higher VIQ was associated with faster RT. These findings suggest that stronger language skills in ASD were uniquely associated with more efficient false belief reasoning, a relationship that was not evident in the TD group.

Despite the correlation between VIQ and false belief RT among participants with ASD in the current study, however, there was no correlation between another measure of language ability – CELF Core Language – and false belief RT. One potential explanation for this inconsistent finding relates to the CELF's lack of sensitivity to differences in language ability. In the current study, for example, the lowest CELF standard score was 93. Moreover, the overall mean score was 112, and the SD was 10.3 (i.e., less than the expected standard score SD of 15); for the VIQ, mean was 10.7 and SD was 2.5, providing a wider range of scores. In contrast to VIQ scores, therefore, the CELF may fail to capture individual differences in language abilities. Thus, it is

possible that the limitations of the CELF scores account for the lack of correlation to false belief performance within the ASD group in this study. Another (somewhat unlikely) possibility is that false belief processing among ASD participants was specifically associated with expressive vocabulary, which was the metric of VIQ in this study, rather than more general language abilities, as captured by the CELF. Regardless, the lack of association between CELF Core language and false belief performance indicates that the VIQ/false belief finding should be interpreted with caution. Given these potentially conflicting results, additional study of the relationship between language and belief reasoning among ASD individuals is warranted.

Nonetheless, the relationship between VIQ and false belief processing within the ASD group calls into question the theory that people with ASD rely *less* on language for mediating ToM tasks. How do we explain this association between language skills and false belief-reasoning skills within our ASD group? First, although some studies have found indirect evidence that people with ASD mediate cognitive tasks via visuospatial systems rather than language, other studies have found that people with ASD *do* use inner speech to mediate cognitive tasks, provided that they have adequate language (Williams, Happe, & Jarrold, 2008). In contrast to the popular idea that people with high functioning ASD "think in pictures" instead of words, our findings and the literature more broadly suggest that this phenomenon may not apply to all people with ASD in all situations.

Still, the current results suggest that linguistic competence has a unique bearing on ToM in ASD, and that people with ASD rely *more* heavily on language to reason about beliefs. This is consistent with previous findings of a stronger relationship between language abilities and false belief task performance in ASD as compared to other populations. Happé (1995) found that, although lexical knowledge was associated with false belief task performance in both TD and

ASD groups, the level of lexical knowledge needed to pass false belief tasks was significantly higher for children with ASD than TD children. Another study found a stronger correlation between mastery of sentence complement syntax and false belief task performance within an ASD group than compared to a TD group (Lind & Bowler, 2009). Likewise, Fisher, Happé, and Dunn (2005) found a stronger association between language skills (especially grammar) and false belief task performance among children with ASD than children with moderate learning difficulties, and Hale and Tager-Flusberg (2003) found that production of sentence complements predicts belief reasoning ability.

How do we explain this unique relationship between language and ToM in ASD? Why, in the current study and others, do people with ASD appear to rely *more* upon language when reasoning about others' beliefs? One possibility relates to the two-systems account of ToM. Typically developing individuals may be able to rely largely upon the intuitive, and automatic "social-perceptive" ToM when solving false belief tasks. Because one of the signature features of ASD is "failure of intuitive mentalizing," people with ASD may solve false belief tasks but succeed only by using the deliberate, effortful, and cognitively taxing "social-cognitive" approach, to reason their way through false belief tasks (Frith, 2004; Tager-Flusberg, 2001). This may explain why people with ASD tend to demonstrate more prominent ToM deficits in naturalistic as compared to laboratory settings (Apperly & Butterfill, 2009). Because of this reliance on the social-cognitive ToM system, people with ASD are especially dependent upon language to facilitate or "bootstrap" their false belief task performance (Joseph, McGrath, & Tager-Flusberg, 2005; Tager-Flusberg, 2001). For example, in the absence of representational, social-perceptive ToM, complement syntax affords a means of "hacking out" solutions to false belief tasks (de Villiers & Pyers, 2002; Lind & Bowler, 2009). In contrast to the theory

suggesting that ToM deficits result from deficits in internalized language among people with ASD, this theory suggests that internalized language, in fact, helps people with ASD compensate for deficits in social-perceptive ToM. As such, successful false belief task performance in ASD may not reflect the same underlying process that operates in TD. Without intuitive ToM, belief reasoning is a cognitively taxing process and draws more heavily on language. This explains why, in the current study, language skills were correlated with false belief processing among ASD participants but not TD participants. Moreover, the current finding that ASD participants were, on average, almost a full second slower than their TD peers on the belief-reasoning task is consistent with a slower, more deliberate "hacking-it-out" approach to belief-reasoning.

In many circumstances (e.g., during particularly executively-demanding tasks), TD individuals also rely heavily on language during certain ToM tasks. The difference, however, is that these TD individuals are also able to draw on intuitive, social-perceptive ToM in the process of solving the task. Indeed, one potential extension of this theory relates to a vicious cycle between language and social deficits: many people with ASD have language deficits at an early age, which, in conjunction with their poor intuitive mentalizing, likely delays their ToM development, which hinders their social skills, which, in turn hinders their language development.[6]

Group difference in RT. In addition to the aforementioned results, we also found a fairly dramatic group difference in RT, where ASD participants were nearly a full second slower than TD participants across conditions. This finding is consistent with other research showing that individuals with ASD tend to have generally slower processing speed (Mayes & Calhoun, 2007) and slower reaction time (Schmitz, Daly, & Murphy, 2007) than their TD peers. It is interesting

[6] One alternative interpretation is that this finding is due to the fact that the TD group had less VIQ variance than the ASD group (TD range = 9-16 (*SD* = 2.0); ASD range = 5-13 (*SD* = 2.7)). It is therefore possible that the lack of an association between VIQ and false belief RT in the TD group was primarily due to this limited range of VIQ scores.

to note that, despite this reaction time difference, there were no group differences in accuracy.

Executive functions and belief reasoning RT. We found no correlation between EF or inhibitory control and false belief RT. This finding is particularly perplexing given the association between VIQ and false belief RT among ASD participants. Considering the theory that belief reasoning is more cognitively taxing among people with ASD, we would predict that EF skills would also be associated with false belief performance in this population. Given the necessarily small size of the ASD group, and given that this was a secondary line of inquiry in the present study, additional research is needed to more fully examine the relationship among language, EF, and belief reasoning in ASD.

Limitations, strengths, and future directions. There were several strengths of the current study. In contrast to many prior dual-task studies, our verbal and spatial interference tasks were comparable in terms of general attentional and cognitive resources, as confirmed by Study 2. Therefore, performance differences between dual task conditions were likely due to language demands on the system rather than general cognitive demands. Moreover, unlike some previous studies, our belief-reasoning task included human actors in realistic settings.

There were also several limitations. First, the experimenter was not naïve to diagnostic condition. Given that the task was computerized, this likely had negligible effects. Nonetheless, future studies would benefit from keeping experimenters blind to diagnostic status when possible. Secondly, both measures of language ability had limitations. As mentioned above, the restricted range of CELF scores may have limited the sensitivity of our correlational analyses. Moreover, although we found a significant correlation between VIQ and false belief processing among ASD participants, these VIQ scores were based upon a single subtest – the Stanford-Binet vocabulary verbal routing subtest – and should therefore be interpreted with caution. Future

studies may consider using more comprehensive measures of VIQ or more subtle online language processing measures (e.g., Eigsti & Bennetto, 2009).

Also, because we did not measure accuracy on the load tasks, we were unable to probe for tradeoffs between load and belief-reasoning tasks. Although, in our subjective assessment, all participants reliably engaged in the shadowing tasks during the belief-reasoning task, it is possible that some participants sacrificed verbal or spatial shadowing performance in favor of true or false belief task performance, which would not be captured in the current results. Additionally, although necessary given our limited sample size, the repeated-measures nature of the experimental design was also a study limitation given the nature of the belief-reasoning task. The repetition of several generally-similar false belief trials, as well as the interweaving of true belief trials, ultimately detracted from the original intention of the location-change false belief task – that is, to test one's capacity for applying ToM in real time. The repetitious nature of the current task effectively detracted from the salience of ToM in solving the task, as participants were able to "get into set" after one or two belief task trials. Moreover, the repeated measures nature of the task rendered the true belief trials a poor "control" condition. That said, results from the first trial only were similar to results from the entire set.

Future investigators should be especially thoughtful about the design and demands of the belief-reasoning task selected. In particular, false belief tasks should be used with caution, especially when conducting research on adolescents and adults, for whom the classic false belief task may have limited utility. Relatedly, given the current lack of clarity about the relationship between internalized language and executive functions between ASD and TD individuals, future studies should also continue to include executive tasks to examine language in relation to ToM task performance. This would allow for more deliberate examination of various executive

functions as potential mediators of language and ToM. Additional research is needed to elucidate the relationship among language ability, internalized language, executive functions, and ToM. As discussed above, more systematic research is also needed to clarify the cognitive processes involved in ToM more generally, and to determine whether ToM is, in and of itself, a useful construct.

Conclusion. To our knowledge, the current study was the first to directly examine the relationship between language and internalized speech and false belief in an ASD population. Our findings, which were discrepant from Newton and de Villiers (2007), highlight broader methodological and theoretical issues within the ToM literature. Our unexpected finding of an association between language and false belief performance not among TD participants, but among ASD participants, supports the account that ToM is not a singular entity, but rather operates in two parallel systems: an automatic and intuitive "social-perceptive ToM," and an effortful and verbally taxing "social-cognitive ToM." Though TD individuals are able to engage both ToM systems, depending upon the given context, people with ASD, who tend to lack intuitive theory of mind, may rely on language to "bootstrap" their ToM. If supported by additional research, this finding would inform our conceptualization of ToM deficits in ASD. In fact, ASD interventions, such as social skills trainings, could yoke this compensatory strategy of slowly but effectively "hacking out" ToM via language. Likewise, this finding highlights the protective nature of language, and therefore, the importance of fostering early language skills among children with ASD. The "social-cognitive ToM" account of ASD certainly warrants further study.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders IV-TR* (IV-TR ed.). Washington, DC: American Psychiatric Association.

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. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/17470218.2012.676055?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953. Retrieved from http://psycnet.apa.org/journals/rev/116/4/953/

Baddeley, A., Chincotta, D., & Adlam, A. (2001). Working memory and the control of action: evidence from task switching. *Journal of Experimental Psychology: General, 130*(4), 641. Retrieved from http://psycnet.apa.org/journals/xge/130/4/641/

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General, 128*(1), 32. Retrieved from http://psycnet.apa.org/journals/xge/128/1/32/

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychological Bulletin, 121*(1), 65-94. Retrieved from http://psycnet.apa.org/journals/bul/121/1/65/

Bialystok, E., & Viswanathan, M. (2009). Components of executive control with advantages for bilingual children in two cultures. *Cognition, 112*(3), 494-500. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2755257/pdf/nihms127345.pdf

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. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027700000962

Carlson, S., Mandell, D., & Williams, L. (2004). Executive function and theory of mind: stability and prediction from ages 2 to 3. *Developmental Psychology, 40*(6), 1105-1122. Retrieved from http://psycnet.apa.org/journals/dev/40/6/1105/

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*. Retrieved from http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=172512

Channon, S., Charman, T., Heap, J., Crawford, S., & Rios, P. (2001). Real-life-type problem-solving in Asperger's syndrome. *Journal of Autism and Developmental Disorders, 31*(5), 461-469.

de Villiers, J. G. (2007). The interface of language and Theory of Mind. *Lingua: International Review of General Linguistics*. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0024384107000113

de Villiers, J. G., & de Villiers, P. A. (2005). Can language acquisition give children a point of view?

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development*.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System (D-KEFS) Examiner's Manual*. San Antonio, TX: The Psychological Corporation.

Dennis, M., Francis, D. J., Cirino, P. T., Schachar, R., Barnes, M. A., & Fletcher, J. M. (2009). Why IQ is not a covariate in cognitive studies of neurodevelopmental disorders. *Journal of the International Neuropsychological Society, 15*, 1-13.

Eigsti, I. M., & Bennetto, L. (2009). Grammaticality judgments in autism: Deviance or delay. *Journal of Child Language, 36*(5), 999-1021. Retrieved from http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=6407080

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests*: Educational Testing Service (ETS).

Fernyhough, C. (1996). The dialogic mind: A dialogic approach to the higher mental functions. *New Ideas in Psychology, 14*(1), 47-62.

Fernyhough, C. (2008). Getting Vygotskian about theory of mind: Mediation, dialogue, and the development of social understanding. *Developmental Review, 28*(2), 225-262.

Fernyhough, C., & Fradley, E. (2005). Private speech on an executive task: Relations with task difficulty and task performance. *Cognitive Development, 20*(1), 103-120.

Fisher, N., Happé, F., & Dunn, J. (2005). The relationship between vocabulary, grammar, and false belief task performance in children with autistic spectrum disorders and children with moderate learning difficulties. *Journal of Child Psychology and Psychiatry, 46*(4), 409-419. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1469-7610.2004.00371.x/abstract

Frith, U. (2004). Emanuel Miller lecture: confusions and controversies about Asperger syndrome. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 45*(4),

672-686. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15056300

Frith, U., Happé, F., & Siddons, F. (1994). Autism and theory of mind in everyday life. *Social Development, 3*, 108–124.

Gergely, G., Nádasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165-193. Retrieved from http://www.sciencedirect.com/science/article/pii/001002779500661H

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). Behavior rating inventory of executive function. *Child Neuropsychology, 6*(3), 235-238. Retrieved from http://www.ncbi.nlm.nih.gov

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: a training study. *Dev Sci, 6*(3), 346-359.

Happé, F. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*(3), 843-855. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7789204

Happé, F., & Frith, U. (1996). The neuropsychology of autism. *Brain, 119*(4), 1377-1400. Retrieved from http://brain.oxfordjournals.org/cgi/doi/10.1093/brain/119.4.1377

Hegarty, M., Shah, P., & Miyake, A. (2000). Constraints on using the dual-task methodology to specify the degree of central executive involvement in cognitive tasks. *Memory & Cognition, 28*(3), 376-385.

James, W. (1890). *The Principles of Psychology* (Vol. 1): Dover Publications; Reprint Edition (June 1, 1950).

Joseph, R. M., McGrath, L. M., & Tager-Flusberg, H. (2005). Executive dysfunction and its relation to language ability in verbal school-age children with autism. *Developmental Neuropsychology, 27*(3), 361-378.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist, 58*(9), 697. Retrieved from http://psycnet.apa.org/journals/amp/58/9/697/

Kaland, N., Callesen, K., Møller-Nielsen, A., Mortensen, E. L., & Smith, L. (2008). Performance of children and adolescents with Asperger syndrome or high-functioning autism on advanced theory of mind tasks. *Journal of Autism and Developmental Disorders, 38*(6), 1112-1123. Retrieved from http://link.springer.com/article/10.1007%2Fs10803-007-0496-8

Kana, R. K., Keller, T. A., Cherkassky, V. L., Minshew, N. J., & Just, M. A. (2006). Sentence comprehension in autism: Thinking in pictures with decreased functional connectivity. *Brain, 129*(9), 2484-2493.

Leslie, A. M. (1994). Pretending and believing: Issues in the theory of ToMM. *Cognition, 50*(1), 211-238.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in cognitive sciences, 8*(12), 528-533. Retrieved from http://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(04)00260-8?_returnURL=http%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1364661304002608%3Fshowall%3Dtrue

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental science, 1*(2), 247-253.

Lidstone, J. S., Meins, E., & Fernyhough, C. (2010). The roles of private speech and inner speech in planning during middle childhood: Evidence from a dual task paradigm. *Journal of Experimental Child Psychology, 107*(4), 438-451. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022096510001128

Lind, S. E., & Bowler, D. M. (2009). Language and theory of mind in autism spectrum disorder: the relationship between complement syntax and false belief task performance. *Journal of Autism and Developmental Disorders, 39*(6), 929-937.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2002). *Autism Diagnostic Observation Schedule (ADOS)*. Los Angeles: Western Psychological Services.

Mayes, S. D., & Calhoun, S. L. (2007). Learning, attention, writing, and processing speed in typical children and children with ADHD, autism, anxiety, depression, and oppositional-defiant disorder. *Child Neuropsychology, 13*(6), 469-493.

Meltzoff, A. N. (1995). Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*(5), 838. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4137788/pdf/nihms459908.pdf

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*(2), 622-646. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8624.2007.01018.x/full

Murray, D. J. (1967). The role of speech responses in short-term memory. *Canadian Journal of Psychology/Revue canadienne de psychologie, 21*(3), 263.

Newton, A. M., & de Villiers, J. G. (2007). Thinking while talking: adults fail nonverbal false-belief reasoning. *Psychological Science, 18*(7), 57.

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development, 67*(2), 659-677.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255-258. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3357322/pdf/nihms375148.pdf

Roid, G. (2003). *Stanford-Binet Intelligence Scales - 5th Edition*. Itasca, IL: Riverside Publishing.

Rutter, M., Bailey, A., & Lord, C. (2003). *The Social Communication Questionnaire (SCQ)*. Los Angeles: Western Psychological Services.

Schmitz, N., Daly, E., & Murphy, D. (2007). Frontal anatomy and reaction time in autism. *Neuroscience Letters, 412*(1), 12-17. Retrieved from http://www.sciencedirect.com/science/article/pii/S0304394006007178

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals* (4th ed.). San Antonio, TX: Harcourt Assessment, Inc.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580-586. Retrieved from http://pss.sagepub.com/content/18/7/580.long

Tager-Flusberg, H. (2001). A re-examination of the theory of mind hypothesis of autism. In J. Burack, T. Charman, N. Yirmiya, & P. R. Zelazo (Eds.), *Development in autism: Perspectives from theory and research* (pp. 173–193). Hillsdale, NJ: Erlbaum.

Vandierendonck, A., Kemps, E., Fastame, M. C., & Szmalec, A. (2004). Working memory components of the Corsi blocks task. *British Journal of Psychology, 95*(1), 57-79.

Vygotsky, L. S. (1987). Thinking and speech *The collected works of LS Vygotsky* (Vol. 1, pp. 39-285).

Wallace, G. L., Silvers, J. A., Martin, A., & Kenworthy, L. E. (2009). Brief report: Further evidence for inner speech deficits in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*(12), 1735-1739. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2974759/pdf/nihms-223889.pdf

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA:: Bradford Books/MIT Press.

Whelan, R. (2010). Effective analysis of reaction time data. *The Psychological Record, 58*(3), 9.

Whitehouse, A. J., Maybery, M. T., & Durkin, K. (2006). Inner speech impairments in autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 47*(8), 857-865.

Williams, D., Bowler, D., & Jarrold, C. (2012). Inner speech is used to mediate short-term memory, but not planning, among intellectually high-functioning adults with autism spectrum disorder. *Development and Psychopathology, 24*(01), 225-239.

Williams, D., Happe, F., & Jarrold, C. (2008). Intact inner speech use in autism spectrum disorder: evidence from a short-term memory task. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 49*(1), 51-58.

Winsler, A., & Naglieri, J. (2003). Overt and covert verbal problem-solving strategies: Developmental trends in use, awareness, and relations with task performance in children aged 5 to 17. *Child Development, 74*(3), 659-678. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/1467-8624.00561/abstract

Woolfe, T., Want, S. C., & Siegal, M. (2002). Signposts to development: theory of mind in deaf children. *Child Development, 73*(3), 768-778. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/1467-8624.00437/abstract