Journal of Autism and Developmental Disorders
https://doi.org/10.1007/s10803-021-05040-6

BRIEF REPORT

Brief Report: Test–Retest Reliability of Cognitive, Affective, and Spontaneous Theory of Mind Tasks Among School-Aged Children with Autism Spectrum Disorder

Melody R. Altschuler[1,2] · Susan Faja[3,4]

Accepted: 20 April 2021

Abstract

The present study evaluates the test–retest reliability of six theory of mind (ToM) tasks that measured cognitive, affective, and spontaneous ToM in 7 to 11 year-old children with autism spectrum disorder. Our results revealed considerable variation in test–retest reliability depending on the type of ToM task, which ranged from poor to good with the majority of the measures exhibiting moderate reliability. Results inform which common measures of cognitive ToM should be selected versus avoided in future intervention work, suggest our measure of spontaneous ToM should be used more widely in intervention and ToM research more broadly, and indicate more work is needed to develop reliable measures of affective ToM. Implications for research and clinical practice are discussed.

Keywords Theory of mind · Test–retest reliability · Social cognition · Autism spectrum disorder · Interventions

Introduction

Theory of mind (ToM) difficulties are central to autism spectrum disorder (ASD), both in terms of worse performance compared to typical development and associations with the dimensional social symptoms associated with ASD (Baron-Cohen et al., 2000). With the goal of ameliorating the social difficulties that many individuals with ASD face, researchers are interested in developing targeted interventions that will improve ToM and thereby enhance associated social and adaptive skills in ASD (e.g., McMahon et al., 2013; Soorya et al., 2015).

A critical feature of a measure used in an intervention is adequate test–retest reliability; however, the research on test–retest reliability of ToM measures in ASD is limited. This is especially critical for tasks that use vignettes or stories to assess understanding of a situation because repetition impacts knowledge of the task. To our knowledge, only a few studies have examined test–retest reliability of ToM measures in individuals with ASD. These include caregiver-questionnaire ToM measures that have revealed excellent test–retest reliability: the Perceptions of Children's Theory of Mind Measure (Hutchins et al. 2008) and the Theory of Mind-Inventory (Hutchins et al., 2012). Other studies have examined test–retest reliability of task-based assessments of ToM in ASD, including the Strange Stories test, a measure of advanced ToM in adults with ASD that has shown fair to good test–retest reliability across the different stories embedded in the task (Shahrivar et al., 2017). The other two task-based assessments of ToM were examined in children, including the ToM Storybooks task, which showed good test–retest reliability in children with Pervasive Developmental Disorder-Not Otherwise Specified (Blijd-Hoogewys et al., 2008), and the Theory of Mind Task Battery, which showed adequate test–retest reliability across the tasks in children with ASD (Hutchins et al., 2008).

✉ Susan Faja
susan.faja@childrens.harvard.edu

Melody R. Altschuler
altsc012@umn.edu

[1] Institute of Child Development, University of Minnesota, Minneapolis, MN, USA

[2] Department of Psychology, University of Minnesota, Minneapolis, MN, USA

[3] Laboratories of Cognitive Neuroscience, Boston Children's Hospital, Two Brookline Place, Brookline, MA 02445, USA

[4] Department of Pediatrics/Division of Developmental Medicine, Department of Psychiatry and Behavioral Sciences, Harvard Medical School, Boston, MA, USA

Published online: 12 May 2021

Beyond the paucity of studies examining the test–retest reliability of ToM in ASD is the issue of distinct theoretical types of ToM with varying levels of complexity, resulting in many different assessments of ToM and its subtypes (Apperly, 2012). The abilities to reason about others' beliefs (cognitive ToM) versus emotions (affective ToM) have been found to be distinct types of ToM, both in typical development (Shamay-Tsoory & Aharon-Peretz, 2007) and in ASD (Altschuler et al., 2018). Spontaneous ToM—the ability to spontaneously identify pertinent social information before utilizing relevant social cognitive skills—is another type with particular relevance for the challenges that individuals with ASD face spontaneously responding to real-life social scenarios (Klin et al., 2003). However, no study to date has examined the test–retest reliability of cognitive, affective, and spontaneous ToM tasks in a single sample.

Present Study

The aim of this study was to evaluate the test–retest reliability of an extensive battery of six ToM tasks that measured cognitive, affective, and spontaneous ToM in 7 to 11 year-old children with ASD. Test–retest reliability was measured via intraclass correlation coefficients (ICCs) for the percent correct obtained for performance on each ToM measure. We hypothesized that the ToM measures would have moderate reliability overall. Moreover, we used Fisher's *r*-to-*z* transformations to compare the strength of ICCs across the four measures of cognitive ToM.

Method

Participants

As part of a larger clinical trial examining the effects of a three-month executive function intervention, children with an existing ASD diagnosis were recruited from a participant registry in a hospital setting and local community sources. Exclusionary criteria included severe sensory or motor impairments that limited the ability to complete the test battery, colorblindness, insufficient English fluency for valid completion of standardized measures, medical disorders that impact the central nervous system, prolonged prenatal substance exposure, and a history of seizures or use of anticonvulsant medications. The hospital's human subject's division approved all study procedures, and all parents consented for their children to participate.

Thirty-five children (4 females) between the ages of 7 to 11 years-old with ASD and with an intelligence quotient (IQ) of 80 or above participated. Descriptive statistics for sample characteristics, including gold-standard measures of ASD, are reported in Table 1. In addition, caregivers reported the following sample characteristics when given a list of options to choose from: primary caregiver's highest education level (high school graduate $n = 2$, associate degree $n = 5$, some college $n = 10$, bachelor degree $n = 16$), annual household income (< 35 K $n = 4$, 36-65 K $n = 5$, 66-100 K $n = 9$, 101-160 K $n = 6$, > 160 K $n = 7$), child's race (Asian $n = 1$, Black/African American $n = 3$, White/Caucasian $n = 27$, more than one race $n = 4$), and Latino/Hispanic yes or no (yes $n = 1$, no $n = 31$).

Table 1 Sample characteristics (N = 35)

	M (SD)	Range
Age (years)	9.10 (1.34)	7.17–11.83
ADOS-2 Comparison Scores		
Total	8.89 (1.23)	6–10
Social affect	8.40 (1.24)	6–10
Restricted repetitive	9.06 (1.31)	4–10
ADI-R Raw Scores		
Social	18.26 (5.16)	10–29
Verbal communication	16.63 (4.54)	8–24
Restricted and repetitive behavior	8.43 (2.32)	3–14
Vineland-2		
Adaptive behavior composite	84.51 (8.69)	69–105
Communication domain	91.00 (10.19)	74–117
Daily living skills domain	87.46 (9.88)	73–107
Socialization domain	80.80 (10.33)	64–112
WASI-2		
Full scale IQ	102.83 (12.15)	80–127
Verbal comprehension index	102.40 (12.40)	74–127
Perceptual reasoning index	102.74 (13.09)	69–127

ADOS-2 Autism Diagnostic Observation Schedule, Second Edition, *ADI-R* Autism Diagnostic Interview, Revised, *Vineland-2* Vineland Scales of Adaptive Behavior, Second Edition, *WASI-2* Wechsler Abbreviated Scale of Intelligence, Second Edition

All participants had a previous ASD diagnosis, which was confirmed at the initial time point using the Autism Diagnostic Observation Schedule, Second edition (ADOS-2; Lord et al., 2012), the Autism Diagnostic Interview-Revised (ADI-R; Rutter et al., 2003) according to Collaborative Programs of Excellence in Autism criteria (see Sung et al., 2005 for details), and the Diagnostic and Statistical Manual of Mental Disorders, Fifth edition (DSM-5; American Psychiatric Association, 2013) criteria for ASD. All children were randomly assigned to the waitlist condition (i.e., no intervention) of the larger clinical trial. Therefore, participants provided data at two time points, approximately three months apart, without receiving the study intervention in that period. Two children were lost to follow-up between time points.

Procedure

Basic exclusion criteria were screened by phone. Then, the Vineland Scales of Adaptive Behavior-2 (Vineland-2; Sparrow et al., 2005), ADI-R, ADOS-2, Wechsler Abbreviated Scale of Intelligence-2 (WASI-2; Wechsler, 2011), and a colorblindness test were completed to determine eligibility. An identical battery of ToM measures was administered at time points 1 and 2.

Materials

Multiple measures were used to assess different types of ToM. All ToM stimuli (pictures, text, audio, and video recordings) and task instructions were presented via computer on E-prime 2.0 software. See Altschuler et al. (2018) for a detailed description of study procedures and ToM tasks, including inter-rater reliability.

In the First-Order False Belief Videos tasks, children answered questions after watching two videos. In the Location Change False Belief task (Saxe, 2009; Wimmer & Perner, 1983), children inferred the knowledge and belief of a person regarding the location of an object that was moved to a new location while he was absent. In the Unexpected Contents False Belief task (Perner et al., 1987), children viewed a familiar container with unexpected contents and were asked about the initial belief of a naïve person regarding the contents of the box. Percent correct was calculated by dividing the number of correct answers to test questions by the total number of test questions for each task (4 for Location Change and 3 for Unexpected Contents).

In the Theory-of-Mind Test (TOM Test; Muris et al., 1999), children answered questions about a series of cartoons and audio stories designed to test both basic and more advanced aspects of ToM. The task contained 38 items and three subscales: TOM Level 1 (20 items), TOM Level 2 (13 items), and TOM Level 3 (5 items). TOM Test Level 1 measures affective ToM, such as recognition of others' affective mental states and understanding of social scenarios and emotions. TOM Test Level 2 measures first-order false belief. TOM Test Level 3 measures second-order false belief. The experimenter scored the child's responses, based on a scoring sheet with common correct and incorrect responses provided by Muris et al. (1999). The responses flagged for review during administration were resolved post-administration by consensus coding and reviewed by the senior author. Percent correct scores were calculated by dividing the total number of correct answers by the total number of questions for each level.

In the Social Attribution Task (SAT; Klin, 2000), children viewed animated geometrical figures that are commonly interpreted as enacting a social scene (Heider & Simmel, 1944). Children were asked to describe the meaning of the animation. The SAT video clips, instructions, and coding scheme were identical to those used by Klin (2000), although the only index coded in the present study was the Problem Solving Index given the high correlation and with the other SAT indices, as reported in previous work (Altschuler et al., 2018), and that this is the least time-consuming to code and therefore most feasible for a measurement of change in social cognition for interventions. All responses were recorded for transcription and coding. To quantify spontaneous ToM, the SAT Problem Solving Index was derived from the spontaneous narratives generated by participants (Klin, 2000). The SAT Problem Solving Index measures the ability to correctly answer questions about the animation. It is scored by summing the number of correct responses (from a total of 10 items) divided by 10, with higher scores representing an increased ability to make salient social attributions when presented with questions about the social nature of the scenes. To maximize inter-rater reliability (Klin, 2000): (1) three coders were trained on SAT scoring before coding the transcripts included in this study, and frequent meetings were held to learn explicit scoring guidelines and to clarify coding issues, (2) coders followed a procedural sequence for coding each transcript, and (3) examples of frequent terms encountered in SAT narratives were included in the manual with their corresponding codes.

Analytic Approach

Analyses were performed in SPSS Version 24.0 (IBM Corp) using a two-way mixed effects model with absolute agreement (Koo & Li, 2016). As specified by standard reliability classification rates (Portney & Watkins, 2009), ICC values < 0.5 indicated poor reliability, ICC values between 0.5 and 0.75 indicated moderate reliability, ICC values between 0.75 and 0.9 indicated good reliability, and ICC values > 0.9 indicated excellent reliability.

Results

Table 2 indicates the number of participants, descriptive statistics, and ICC for each ToM measure at time points 1 and 2. Measures of first-order false belief ranged from poor to moderate reliability: TOM Test Level 2 had poor reliability, whereas Unexpected Contents had moderate reliability, and Change in Location had moderate reliability. The measure of second-order false belief from TOM Test Level 3 had moderate reliability. In contrast, the measure of affective ToM from TOM Test Level 1 had poor reliability. The highest reliability was for the measure of spontaneous ToM from the SAT Problem Solving Index, which had good reliability.

Fisher's r-to-z transformations revealed that across all four measures of cognitive ToM, the two measures that

Table 2 Descriptive statistics and ICC estimates for ToM measures

	Time 1		Time 2		Single measure ICC [95% CI]
	Mean (SD)	*n*	Mean (SD)	*n*	
Cognitive ToM: FOFB					
TOM test level 2	0.55 (0.17)	35	0.59 (0.14)	33	0.33 [−0.007 to 0.60]*
Unexpected contents	0.66 (0.32)	33	0.54 (0.38)	32	0.50 [0.18–0.73]*
Change in location	0.74 (0.36)	34	0.83 (0.29)	32	0.55 [0.26–0.75]**
Cognitive ToM: SOFB					
TOM test level 3	0.38 (0.29)	35	0.44 (0.34)	33	0.72 [0.51–0.85]**
Affective ToM					
TOM test level 1	0.73 (0.14)	35	0.75 (0.12)	33	0.49 [0.18–0.71]*
Spontaneous ToM					
SAT problem solving index	0.30 (0.19)	27	0.28 (0.16)	26	0.78 [0.56–0.90]**

ICC intraclass correlation coefficients, *ToM* theory of mind, *FOFB* first-order false belief, *SOFB* second-order false belief

*$p < 0.01$

**$p < 0.001$

significantly differed from each other were TOM Test Level 2 and TOM Test Level 3 ($z = -2.19$, $p = 0.01$).

We next explored age-related differences in test–retest reliability (see Table 3). To do so, we used a median split of 9 years of age (at time point 1), to differentiate our sample into younger and older groups. Using this criterion, 17 children (15 males, $M_{age} = 7.97$, $SD = 0.58$ years) were grouped into the younger cohort (7.17 to 8.83 years) and 18 children (16 males, $M_{age} = 10.16$, $SD = 0.92$ years) were grouped into the older cohort (9 to 11.83 years). Fisher's *r*-to-*z* transformations revealed that across all measures of ToM, the test–retest reliability of the younger and older groups did not differ from each other.

Discussion

The aim of the current study was to evaluate test–retest reliability of a large battery of six ToM measures that assessed cognitive, affective, and spontaneous ToM in children with ASD between the ages of 7 to 11 years-old. Only a handful of studies have examined test–retest reliability of ToM measures in children with ASD to date. These studies have revealed mixed results, ranging from poor to excellent test–retest reliability (Blijd-Hoogewys et al., 2008; Hutchins et al., 2012; Hutchins et al., 2008; Hutchins et al., 2008; Shahrivar et al., 2017), and none of them have examined test–retest reliability using an extensive battery of ToM measures that span the domains of cognitive, affective, and

Table 3 ICC estimates and Fisher's *r*-to-*z* for ToM measures by younger versus older age groups

	Younger age cohort	Older age cohort	Fisher's *r*-to-*z* (*p*-value)
	Single measure ICC [95% CI]	Single measure ICC [95% CI]	
Cognitive ToM: FOFB			
TOM test level 2	0.57 [0.14–0.82]*	0.09 [− 0.43 to 0.54]	1.5 (0.07)
Unexpected contents	0.52 [0.04–0.81]*	0.49 [− 0.02 to 0.79]*	0.11 (0.46)
Change in location	0.59 [0.09–0.85]*	0.43 [− 0.01 to 0.74]*	0.59 (0.28)
Cognitive ToM: SOFB			
TOM test level 3	0.70 [0.34–0.89]**	0.75 [0.44–0.90]**	− 0.28 (0.39)
Affective ToM			
TOM test level 1	0.53 [0.05–0.81]*	0.41 [− 0.07 to 0.73]	0.42 (0.34)
Spontaneous ToM			
SAT problem solving index	0.64 [0.11–0.89]*	0.81 [0.49–0.94]**	− 0.99 (0.16)

ICC intraclass correlation coefficients, *ToM* theory of mind, *FOFB* first-order false belief, *SOFB* second-order false belief

$^{*}p < 0.01$

$^{**}p < 0.001$

spontaneous ToM. Our results indicated that there was considerable variation in test–retest reliability depending on the type of ToM task, which ranged from poor to good with the majority of the measures exhibiting moderate reliability. Specifically, poor test–retest reliability was shown for affective ToM (TOM Test Level 1) and first-order false belief (TOM Test Level 2) when presented as vignettes. Moderate test–retest reliability was shown for first-order false belief videos (Unexpected Contents First-Order False Belief task and Change in Location First-Order False Belief task) and second-order false belief (TOM Test Level 3) vignettes. Finally, good test–retest reliability was shown for the measure of spontaneous ToM (SAT Problem Solving Index).

Interestingly, the measure of ToM—the SAT Problem Solving Index—that showed the best test–retest reliability is also the measure that is rarely used in the literature and has not yet been used in interventions. One likely reason that the SAT is not often used in the literature is that coding of the spontaneous SAT narratives elicited from the film of shapes enacting a social scene is complex and time-consuming to implement and learn, which is a main reason there has been a multiple choice version of the SAT developed (Burger-Caplan et al., 2016). However, the multiple choice version does not elicit spontaneous narratives about the social scene enacted in the SAT, and it is therefore preferable to develop a reliable, yet feasible measure of spontaneous ToM in children with ASD. In a previous examination of the associations between ToM and social symptom severity (Altschuler et al., 2018), we expanded and streamlined the SAT coding system created by Klin (2000) by creating a detailed manual that captured the unique responses of school-aged children with ASD. In the present study, we selected the Problem Solving Index to code given that it is the only SAT index that can be coded in isolation, it is the quickest SAT index to code, and it was correlated with the other indices of spontaneous social attribution at time point 1. The good test–retest reliability of the SAT Problem Solving Index indicates future interventions would benefit from the inclusion of the SAT and our newly refined coding scheme to test whether interventions improve spontaneous ToM and corresponding social functioning.

Another notable finding is that the affective ToM measure did not show strong test–retest reliability, despite the fact that our past work in the same sample suggests affective ToM is uniquely predictive of social symptom severity in ASD (Altschuler et al., 2018). This suggests that work is needed to develop more reliable measures of affective ToM in school-aged children with ASD, as it may be a clinically important domain of ToM to target in interventions that have the goal of ultimately reducing ASD social symptom severity.

Taken together, our results showed variability in test–retest reliability of our large battery of ToM tasks, ranging from poor to good. Poor test–retest reliability was shown for a measure of affective ToM and a measure of first-order false belief vignettes, moderate test–retest reliability was shown for a video measure of first-order false belief and a measure of second-order false belief, and good test–retest reliability was shown for a measure of spontaneous ToM. Together these findings indicate that more work is needed to develop reliable measures of affective ToM, spontaneous ToM can be measured reliably and provide a meaningful way to understand how children with ASD understand social aspects of their environment, and cognitive ToM is more reliably measured with videos of first-order false belief than vignettes. The subgroup analyses by older and younger children revealed a similar pattern of findings and no significant differences between age groups. However, given our small sample size, future work is needed to replicate our results with larger sample sizes and test the possibility that there may be differences in ToM test–retest reliability across development. Overall, these findings highlight the need for rigorous measures of ToM in future efforts that aim to improve social cognitive skills and reduce social difficulties in children with ASD. This work will have important implications both for psychometric and intervention research as well as clinical research and practice more broadly.

Author Contributions Both authors contributed to the study conception and design. Data analysis was performed by SF. The first draft of the manuscript was written by MA and both authors revised previous versions of the manuscript. Both authors read and approved the final manuscript.

Funding Funding for this study was provided by the National Institute of Child Health & Human Development of the National Institutes of Health under Award Number K99/R00HD071966. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program (NSF-GRF 00074041) to M. Altschuler. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank Dr. Ami Klin for helping us learn the SAT coding scheme, the staff and students who assisted with collecting and scoring these measures, and to particularly thank the children and families who participated in this study.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from parents of all individual participants included in the study and all children provided their assent.

References

Altschuler, M., Sideridis, G., Kala, S., Warshawsky, M., Gilbert, R., Carroll, D., Burger-Caplan, R., & Faja, S. (2018). Measuring individual differences in cognitive, affective, and spontaneous theory of mind among school-aged children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 48*(11), 3945–3957

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*. (5th ed.). 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

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. J. (2000). *Understanding other minds: Perspectives from developmental cognitive neuroscience*. Oxford University Press.

Blijd-Hoogewys, E. M. A., Van Geert, P. L. C., Serra, M., & Minderaa, R. B. (2008). Measuring theory of mind in children. Psychometric properties of the ToM storybooks. *Journal of Autism and Developmental Disorders, 38*(10), 1907–1930

Burger-Caplan, R., Saulnier, C., Jones, W., & Klin, A. (2016). Predicting social and communicative ability in school-age children with autism spectrum disorder: A pilot study of the Social Attribution Task Multiple Choice. *Autism, 20*(8), 952–962

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*(2), 243–259

Hutchins, T. L., Bonazinga, L. A., Prelock, P. A., & Taylor, R. S. (2008). Beyond false beliefs: The development and psychometric evaluation of the Perceptions of Children's Theory of Mind Measure—Experimental Version (PCToMM-E). *Journal of Autism and Developmental Disorders, 38*(1), 143–155

Hutchins, T. L., Prelock, P. A., & Bonazinga, L. (2012). Psychometric evaluation of the Theory of Mind Inventory (ToMI): A study of typically developing children and children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 42*(3), 327–341

Hutchins, T. L., Prelock, P. A., & Chace, W. (2008). Test-retest reliability of a theory of mind task battery for children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities, 23*(4), 195–206

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The social attribution task. *Journal of Child Psychology and Psychiatry, 41*(7), 831–846

Klin, A., Jones, W., Schultz, R., & Volkmar, F. (2003). The enactive mind, or from actions to cognition: Lessons from autism. *Philosophical Transactions of the Royal Society of London B, 358*(1430), 345–360

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule*. (2nd ed.). Western Psychological Services.

McMahon, C. M., Lerner, M. D., & Britton, N. (2013). Group-based social skills interventions for adolescents with higher-functioning autism spectrum disorder: A review and looking to the future. *Adolescent Health, Medicine and Therapeutics, 4*, 23

Muris, P., Steerneman, P., Meesters, C., Merckelbach, H., Horselenberg, R., van den Hogen, T., et al. (1999). The TOM test: A new instrument for assessing theory of mind in normal children and children with pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 29*(1), 67–80

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology, 5*(2), 125–137

Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: Applications to practice*. (Vol. 892)Pearson/Prentice Hall.

Rutter, M., Le Couteur, A., & Lord, C. (2003). *Autism diagnostic interview-revised*. Western Psychological Services.

Saxe, R. (2009). How we read each other's minds [Video file]. Retrieved from https://www.ted.com/talks/rebecca_saxe_how_brains_make_moral_judgments

Shahrivar, Z., Tehrani-Doost, M., Khorrami Banaraki, A., Mohammadzadeh, A., & Happe, F. (2017). Normative data and psychometric properties of a Farsi translation of the strange stories test. *Autism Research, 10*(12), 1960–1967

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia, 45*(13), 3054–3067

Soorya, L. V., Siper, P. M., Beck, T., Soffes, S., Halpern, D., Gorenstein, M., et al. (2015). Randomized comparative trial of a social cognitive skills group for children with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*(3), 208–216

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland II: A revision of the vineland adaptive behavior scales: I. Survey/Caregiver form*. American Guidance Service.

Sung, Y. J., Dawson, G., Munson, J., Estes, A., Schellenberg, G. D., & Wijsman, E. M. (2005). Genetic investigation of quantitative traits related to autism: Use of multivariate polygenic models with ascertainment adjustment. *American Journal of Human Genetics, 76*(1), 68–81

Wechsler, D. (2011). *Wechsler abbreviated scale of intelligence*. (2nd ed.). Pearson Assesements.

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.