Focus on Autism and Other Developmental Disabilities
Volume 23 Number 4
December 2008 195-206

10.1177/1088357608322998
http://focus.sagepub.com
hosted at
http://online.sagepub.com

Test-Retest Reliability of a Theory of Mind Task Battery for Children With Autism Spectrum Disorders

Tiffany L. Hutchins
Patricia A. Prelock
University of Vermont, Burlington
Wendy Chace
Franklin County Home Health, St. Albans, Vermont

This study examined for the first time the test-retest reliability of theory-of-mind tasks when administered to children with Autism Spectrum Disorders (ASD). A total of 16 questions within 9 tasks targeting a range of content and complexity were administered at 2 times to 17 children with ASD. In all, 13 questions demonstrated adequate test-retest reliability and high internal consistency. Items that did not achieve reliability violated a pragmatic convention, were ambiguous, or were associated with a response bias. No effect of verbal ability or diagnosis was found on consistency of performance. There was no effect of interval (i.e., short vs. long) on change in score although modest increases in performance occurred generally across administrations. Implications for research and practice are considered.

Keywords: *autism; theory of mind; false belief; assessment; reliability; socialization*

Although directions of influence are difficult to clarify, many researchers have concluded that theory-of-mind (ToM) deficits underlie the social, behavioral, and communicative impairments characteristic of Autism Spectrum Disorders (ASD; e.g., Baron-Cohen, Leslie, & Frith, 1985; Leslie, 1987), making ToM relevant and often central to the study of ASD. Across hundreds of studies, by far the most common ToM assessment strategy makes use of the now classic false belief task (Wellman, Cross, & Watson, 2001; Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998) originally developed by Wimmer and Perner (1983). In this task children are told a story in which an object is moved from an old location to a new location without the knowledge of the main protagonist. For example, Anthony puts a book on the living room table and leaves the room. In his absence, Mariam enters and moves the book from the table to a drawer and then she leaves. Children are asked, "When Anthony returns, where will he look for the book?" Children who answer with the new (incorrect) location fail the question whereas children who answer with the old (correct) location pass the question by presumably demonstrating their knowledge that behaviors are guided by inner mental states, in this case a false belief.

Individuals with ASD generally perform poorly on false belief tasks compared with their age- and language-matched peers. In fact, decades of previous research have demonstrated autism-specific deficits in a range of tasks designed to assess various aspects of logical or belief inferencing. For example, individuals with ASD often experience difficulty during tasks designed to tap the understanding that seeing leads to knowing (e.g., Perner, Frith, Leslie, & Leekam, 1989) and tasks that assess the understanding that people can see the same object in different ways, depending on positioning (e.g., Flavell, Everett, Croft, & Flavell, 1981; Hadwin, Baron-Cohen, Howlin, & Hill, 1996). Not only have deficits been observed for logic and belief inferencing, but individuals with ASD have demonstrated profound deficits in social and emotional perspective-taking skills across a range of targets. For instance, children with ASD often evidence difficulty with emotion recognition (e.g., Prior, Dahlstrom, & Squires, 1990) and the understanding of desire-based emotion (e.g., understanding that people are happy when they get what they want) and belief-based emotion (e.g., understanding that people will be happy if

Authors' Note: Please address correspondence to Tiffany L. Hutchins, 406 Pomeroy Hall, 489 Main Street, University of Vermont, Department of Communication Sciences, Burlington, VT 05405 (e-mail: Tiffany.Hutchins@uvm.edu).

they think they will get what they want; see Baron-Cohen, 1991).

ToM assessment traditionally has relied heavily on the classic false belief task (Wimmer & Perner, 1983), but as ToM is understood to be a complex and multifaceted construct, it is clear that a single, dichotomous measure is inadequate. For this reason, several researchers have argued for the value of aggregate measures in the form of task batteries that assess various aspects of ToM across levels of complexity (e.g., Tager-Flusberg, 2001; Wellman et al., 2001) and that tap related but distinct social cognitive understandings such as those described above (e.g., Hadwin et al., 1996; Hughes et al., 2000). Many batteries conceptualize the content of items as varying along a dimension from simple to complex (e.g., Hadwin et al., 1996; Muris et al., 1999). Assessment of complex or higher level ToM competence typically makes use of second-order false beliefs (e.g., what Wendy thinks Patty thinks) for addressing inferences of both belief and emotion (Hughes et al., 2000; Silliman et al., 2003; Sullivan, Zaitchik, & Tager-Flusberg, 1994).

It is important to note that there has been tremendous variation in the administration of ToM tasks. ToM tasks have been presented using dolls or puppets, storybooks, live actors, and video clips of live actors (e.g., Hadwin et al., 1996; Happe, 1995; Mayes, Klin, Tercyak, Cicchetti, & Cohen, 1996). Wording of the test questions also has differed, as has the order and placement of control questions. These variations may affect performance. For example, Siegal and Beattie (1991) found that adding the word *first* to the standard false belief test question (e.g., "Where will Anthony look for the book first?") elicited more accurate responses, presumably by making clear to the child that the statement should not be interpreted as "Where will Anthony have to look for the book in order to find it?" Similarly, the tendency for children with ASD to interpret language literally may contribute to confusion when interpreting standard tests of false belief. To avoid the potential confusion, Mitchell, Saltmarsh, and Russell (1997) proposed a "message-desire discrepant task," in which a false belief is recast in relation to a desire (e.g., "Which book does Anthony want?"). "The question ostensibly is not about belief but desire, yet the only way of inferring that is by taking into account the speaker's belief and reinterpreting [the] message accordingly—which is to interpret it nonliterally" (p. 689). Clearly, the content, wording, and types of stimuli used when administering ToM tasks have the potential to dramatically influence performance. Careful consideration of these variables is critical to ensure that the tasks are appropriate and sensitive to the challenges faced by individuals with ASD.

Previous Study of the Psychometric Properties of ToM Tasks

A comprehensive review of the literature (including Cochrane Review, ERIC, and PsycINFO) generated only four studies that have specifically examined the test-retest reliability of ToM tasks. The relevant aspects of these studies are summarized in Table 1.

Mayes et al. (1996) were the first to examine the test-retest reliability of false belief tasks in a sample of 23 typically developing children (age 36–71 months). Children were divided into two groups: 11 children were younger than 48 months, and 12 children were older than 48 months. A video format was used to show three false belief tasks, which varied slightly in content and method of administration and incorporated a total of nine test questions. The interval between test and retest was between 2 and 3 weeks. Using kappa, which is a chance-correcting measure of agreement, and a criterion level of at least .40, Mayes et al. reported "poor to moderate" (p. 318) reliability for ToM tasks, with children in the two age groups demonstrating reliable performance in only 4 of 18 comparisons.

Hughes et al. (2000) responded with a test-retest study, over a 4-week interval, on 47 typically developing children ranging in age from 55 months to 62 months. Hughes et al. used puppets and storybooks to present the ToM tasks. Six tasks, designed to represent a range of difficulty, were presented and included a total of nine questions: five first-order false belief questions (e.g., "What does Charlie think is in the box?"), two emotion-inference questions (e.g., "How does Larry feel when he gets a can of Coke?"), and two second-order false belief prediction questions (e.g., "Where does Simon think Mary will look for the chocolate?"). Using kappa, authors reported fair to moderate agreement for eight of the nine test questions. Furthermore, test-retest reliability remained high after partialling the effects of children's verbal mental age (VMA), suggesting no effect of child verbal ability on reliability. Hughes et al. proposed that the poor results found by Mayes et al. (1996) may have been influenced by Mayes et al.'s use of a video format, which is unusual and "may have increased the nonspecific demands of the tasks" (p. 483). Hughes et al. also raised doubts about the scoring procedures used by Mayes et al., which allowed children to pass a false belief question without responding correctly to the relevant memory control questions. Finally, Hughes et al. found high estimates of internal consistency ($\alpha = .84$) and concluded that "the various test questions measure a unitary construct" (p. 488). It should be noted that although indices of internal consistency

Table 1
Previous Investigations of the Test-Retest Reliability of Theory-of-Mind (ToM) Tasks

Study	Participants	Tasks	Test-Retest (Interval, Analysis, Results)	Reliability as a Function of Children's Contrasting Abilities
Charman & Campbell (1997)	36 children, adolescents, and adults with learning disability	Five tasks acted out using dolls and incorporating seven test questions • Three false belief tasks • Two belief–desire reasoning tasks	*Interval* = test and retest conducted in a single session *Analysis*: kappa *Results*: Only three of seven comparisons showed moderate to good agreement.	By the use of VMA scores from the *Test of Reception of Grammar* (Bishop, 1989), reliable passers were found to have higher VMA than did nonreliable passers.
Hughes et al. (2000)	47 typically developing children, 55 to 62 months	Six tasks acted out using dolls and incorporating nine test questions: • Two false belief tasks • Two belief–desire reasoning tasks • Two second-order false belief tasks	*Interval* = 17-39 days (M = 26) *Analysis*: kappa *Results*: Eight of nine comparisons showed fair to moderate agreement.	When scaled IQ scores from vocabulary subtests of the *Wechsler Intelligence Scales for Preschool-* and *Primary-aged Children–Revised* (Wechsler, 1990) were used, no difference in reliability was found when effects of verbal ability were statistically partialled.
Mayes et al. (1996)	23 typically developing children, 36 to 71 months	Three false belief tasks, incorporating nine questions administered using videotapes of live actors	*Interval* = 2-3 weeks *Analysis*: kappa *Results*: Only 4 of 18 comparisons met reliability criteria, which indicates poor to moderate agreement overall.	Children were divided into two groups by age (i.e., less than 48 months and greater than 48 months) and no difference in reliability was found between age groups.
Muris et al. (1999)	12 typically developing children, 5 to 12 years	78 items using either static visual stimuli, interview format, or live actors. Items reflect one of three subscales: • Precursors to ToM (e.g., emotion recognition) • Manifestations of ToM (e.g., false belief tasks) • Advanced ToM (e.g., second-order false belief tasks)	*Interval* = 8 weeks *Analysis*: intraclass correlation coefficient *Results*: coefficient for overall test = .99, indicating sufficient reliability	Not addressed

Note: VMA = verbal mental age.

(e.g., Cronbach's or coefficient alpha) are often interpreted as reflecting the degree to which items on a measure tap a unitary construct, α can be high when constructs are multidimensional. This happens when the different dimensions assessed intercorrelate highly (Nunnally & Berstein, 1994).

In an effort to validate a 78-item ToM test utilizing vignettes, stories, and static visual stimuli, Muris et al. (1999) examined the test-retest reliability of their measure for 12 typically developing children (ages 5–12). The test incorporated an interview format, included a series of *wh-* questions to assess ToM competence, and as such may be appropriate only for typically developing children and older children with ASD who evidence good verbal abilities. The test-retest interval was 8 weeks, and the intraclass correlation coefficient was .99 for the overall test. Because Muris et al. did not specify that their intraclass correlation was calculated to assess absolute agreement, it is likely that it assessed consistency of score. This may have inflated the estimate of temporal stability because this procedure examines the relations between scores but does not penalize for systematic error (e.g., practice effects) and does not correct for chance agreement. In fact, a t test troublingly revealed significant improvements (i.e., absolute differences) between test administrations. Nonetheless, Muris et al. concluded that the test has "sufficient test-retest reliability" (p. 73).

Charman and Campbell (1997) examined the test-retest reliability of three false belief tasks and two belief–desire reasoning tasks for a sample of 36 individuals with a learning disability. Tasks were acted out using dolls, incorporated a total of seven test questions, and unlike most examinations of test-retest reliability, the two administrations of all

tasks were conducted in the same session. Following the statistical (i.e., kappa) and criteria levels adopted by Mayes et al. (1996), Charman and Campbell reported only "moderate" (p. 728) agreement, with three of the seven test questions reaching the criteria for acceptability. It is interesting that reliable passers had higher VMA scores than did unreliable passers.

Statement of the Problem

The few studies examining the test-retest reliability of ToM tasks yielded mixed results, indicating poor to moderate to good reliability. Results also were mixed as to whether children's contrasting ability levels (typically operationalized by VMA) influenced the reliability of ToM tasks. Of course, variation in the results is, in part, a result of variation in the tasks (the content as well as the method of administration), the statistical procedures used to calculate reliability, the nature of the populations sampled, and the length of the test-retest interval. In fact, an important limitation of previous investigations of test-retest reliability of ToM tasks involves examination of a single and relatively short interval. Thus, "the question of longer term test-retest reliability remains an open one for further research" (Wellman et al., 2001, p. 680).

No study to date has specifically examined the test-retest reliability of ToM tasks when used to assess children with ASD, and researchers have cited the need for further work investigating the reliability of ToM tasks in general (Charman & Campbell, 1997; Wellman et al., 2001) and with children with ASD in particular (Mayes et al., 1996). Meanwhile, ToM tasks remain the most popular method for assessing ToM knowledge and thus play a pivotal role in ToM research (Tager-Flusberg, 2001; Wellman et al., 2001; Yirmiya et al., 1998). This is disconcerting because a number of factors likely affect the reliability in performance of children with ASD. Factors that may negatively affect reliability in ToM task performance of children with ASD include, but are not limited to, fluctuations in anxiety, motivation, attention, perseveration, and behavior (Mayes et al., 1996). As such, it is clear that questions involving the psychometric soundness of ToM tasks for this population demand careful examination. Therefore, the following questions were addressed:

- What is the test-retest reliability and internal consistency of a ToM task battery designed to support the performance of children with ASD?
- Does consistency of performance differ between relatively short (2–7 weeks) and longer test-retest intervals (8–16 weeks)?
- Does consistency of performance vary as a function of children's contrasting verbal abilities and diagnoses?

Method

Participants

Participants were 17 children (2 girls, 15 boys) ranging in age from approximately 4.5 years to 12 years (M = 8.3) diagnosed with Autism, Pervasive Developmental Disorder–Not Otherwise Specified (PDD-NOS) or Asperger's Disorder using the criteria in the *Diagnostic and Statistical Manual of Mental Disorders–Fourth Edition* (DSM-IV; American Psychiatric Association, 1994). Diagnoses were made between 22 months and 8 years of age by a developmental pediatrician, pediatric psychiatrist, or psychologist with experience in the diagnosis of children with autism. As stated above, one goal of this study was to examine whether reliability of performance was associated with verbal ability and diagnosis, necessitating variation in the sample in these variables. Eight children had an original diagnosis of autism, seven had a diagnosis of PDD-NOS, and two had a diagnosis of Asperger's Disorder. The term *ASD* is used throughout this article to refer to children with a diagnosis of Autism, PDD-NOS, or Asperger's Disorder. The *Autism Diagnostic Observation Schedule* (ADOS; Lord, Rutter, DiLavore, & Risi, 1999) was used to confirm the participants' diagnoses of ASD. The ADOS was administered by two graduate students in communication sciences who were trained in the administration and scoring of the ADOS by an expert in ASD. All administrations of the ADOS were videotaped and reviewed by the expert to gather estimates of interobserver reliability, which, on average, was 90% (SD = 6; range = 76%–96%) using point-to-point agreement. The diagnosis of ASD was confirmed by the ADOS for all children.

Children represented a range of verbal abilities assessed on the basis of case history and the *Peabody Picture Vocabulary Test–Third Edition* (PPVT-III; Dunn & Dunn, 1997). Although the PPVT-III has demonstrated good psychometric validation (Williams & Wang, 1997), these estimates are for typically developing children. Furthermore, VMA is a developmental score that should not be taken to indicate equivalent performance. As such, scores on this measure should be interpreted with caution. With this in mind, the average PPVT-III score for VMA was 6.7 (range = <2 years–15.25 years). More specifically, seven children obtained PPVT-III scores indicating age-appropriate levels, and two children scored one standard deviation above the mean. Of the remaining eight

children, three obtained PPVT-III scores falling at least one standard deviation below the mean, two obtained scores falling at least two standard deviations below the mean, and three were identified as functionally nonverbal and obtained scores falling at least three standard deviations below the mean. These three children evidenced the lowest VMAs and were between 1.75 years and 3 years old (M = 2.6). The validity of the responses obtained from children with the most limited language abilities is considered in the Discussion section.

ToM Task Battery

A total of nine tasks incorporating 16 test questions were borrowed from tasks developed previously for use with children with ASD. (A complete description of the tasks is available from the first author.) Seven of the nine tasks were borrowed from Hadwin et al. (1996), who developed their items to span a range of content and complexity (i.e., from simple to difficult) although the difficulty of the items has not been empirically evaluated. Items on this test were designed to assess the understanding of emotion- or belief-inferencing during structured ToM teaching episodes. Activities for assessment and teaching utilized a variety of methods, including play, pictures, and games. For the purposes of this study, the content of the ToM assessments developed by Hadwin et al. was adapted in several ways.

Hadwin et al. (1996) varied the number of response options across tasks. For this study, it was often necessary to add response options to ensure that for all tasks, children were presented with one correct response option and three plausible distracters, making the chance of correct responding in the absence of ToM knowledge equal to 25%. This was true for both test and control questions. Some content was altered to ensure that the language used was informal and easily understood. Thus, high-frequency words (e.g., *scared*, *mad*) were substituted for lower-frequency words (e.g., *afraid*, *angry*). Content was also altered to make references to beliefs and emotions more concrete and less syntactically complex (e.g., changing "wanting to catch a fish" to "wanting a cookie"). Because Hadwin et al. combined their ToM assessments with exercises designed to teach ToM, it was also necessary at times to alter the order and placement of control questions. Following the recommendation of Siegal and Beattie (1991), the test question for the item modeled after the classic false belief task was modified to include the word *first* (i.e., "Where will Anthony look for the book first?") to limit the potential that this question would be misinterpreted. Finally, the names of characters in the stories were changed to suit an American sample (e.g., Anthony, Carlos, and Sonya were used in lieu of Katie, Thomas, and Bill). To these seven tasks, an additional two tasks were added: the message–desire discrepant task (Mitchell et al., 1997) and a second-order false belief task (Silliman et al., 2003, adapted from Sullivan et al., 1994). As previously discussed, the message–desire discrepant task was specifically designed to reduce the potential for children with ASD to misinterpret a belief task by recasting a belief in relation to a desire. The second-order false belief task (requiring logical inferencing) was included to ensure that the task battery incorporated what are considered higher order or more-advanced tests of ToM.

Specifically, the first four tasks (1a–1d) targeted the ability to recognize different emotions (i.e., happy, sad, mad, and scared). In the second task (2), participants were asked to infer an emotion based on a situation, and in the third task (3), they were asked to infer an emotion based on a desire. The fourth task assessed more-advanced ability to infer an emotion in three situations: desire–belief (4a), desire–reality (4b), and second-order desire–belief (4c). The remaining five tasks targeted the ability for visual perspective taking and the ability to infer beliefs and actions based on perceptions. The fifth task (5a, 5b) was a line-of-sight visual perspective task. The sixth task asked participants to infer both belief (6a) and action (6b) based on perception (i.e., seeing leads to knowing and behaving). The seventh task (7) was the standard location change, false belief task. The eighth task (8) was the message–desire discrepant task, and the ninth task (9) was a second-order false belief task. Thus, the task battery included simple and more-complex tasks tapping affective and social domains as well as perceptual and logical domains.

In the design of the ToM task battery, care was taken to ensure that the tasks would be sensitive to the challenges faced by verbal and nonverbal children with ASD. For example, task narratives and questions were presented in the form of static visual information in a storybook format to take advantage of the visual learning style that has been identified as a relative strength for many children with ASD (Beukelman & Mirenda, 1992; Dettmer, Simpson, Myles, & Ganz, 2000; Grandin, 1995; Johnston, Nelson, Evans, & Palazolo, 2003). Many pictures also made use of thinking bubbles to make the contents of others' minds explicit through a visual mode. On each page, a limited number of illustrations were presented to avoid the potential detrimental influence of sensory distraction. The text accompanying each illustration (including text for the memory control

and test questions) also appeared on the bottom of each page to ensure the uniformity of the administration procedures and allow children to read along if they were able to and interested in doing so. Children could respond by answering verbally or by pointing to a picture that showed the correct answer. Although seldom used, specific prompts were also included to provide sufficient opportunities for children to respond. In practice, these were used most often for children who were inattentive or unresponsive during testing. Initial prompts requested the children to point to corresponding pictures. If needed, a second prompt verbally repeated the four choices. Each ToM question was scored as pass (1) or fail (0), with a possible total score of 16. In line with previous research, it was not possible for a child to receive credit for a ToM target question without passing the associated control questions.

Procedure

Children were recruited through contacts with school-based speech language pathologists throughout northern and central Vermont to participate in a larger study examining the effects of a Social Story™ intervention for children with ASD. The ToM task battery was administered at two times before the intervention phase of the larger study began. The optimal interval in examinations of test-retest reliability, to minimize complicating influences of maturation, is considered to be less than 1 month for many measures (e.g., McCauley, 2001). On the other hand, potential memory and practice effects can influence performance when the interval is too short (Kaplan & Saccuzzo, 1989). ToM understanding in a sample of children with ASD was not expected to undergo great developmental change, and variation in the interval needed to be sufficiently large to allow for examination of short as well as longer lags. In light of the concerns posed by practice and maturation and the need to examine longer lags (Wellman et al., 2001) a test-retest interval between 2 and 16 weeks was deemed appropriate (range = 20–107 days, $M = 58$, $SD = 28$).

Because familiarity with the setting has been cited as an important factor in the performance of children with ASD, testing was administered on two occasions to each participant in the participant's own home. A different graduate student in communication sciences administered the task at Time 1 and Time 2. In this way, the evaluator was unfamiliar to the child at both administrations, and the evaluator at the second session was blind to the child's performance during the first session. A total of 10 evaluators administered the ToM task batteries to the participants across the two sessions. No effort was made to counterbalance the order of administrators. However, all test givers underwent training, across several sessions, that focused on strategies for successfully engaging children with ASD (including the option of using a visual schedule that was designed to support successful engagement) and instruction that stressed a uniform method of task presentation and prompting through the use of scripted administration procedures and strict scoring protocols. Test administrators were also instructed never to implicitly or explicitly indicate whether a response was correct or incorrect and to avoid verbal praise such as "Good job." They were, instead, instructed to comment solely on the child's engagement in the task (e.g., "I really like how you are listening" or "You're doing such a good job paying attention").

To ensure the uniformity of scoring procedures, 50% of all test administration procedures were videotaped and reviewed by a primary investigator who is also an expert in ASD. Few questions about scoring occurred, and these were discussed and resolved by the group. In particular, it was decided that children who pointed to more than one response option could be asked the question for a second or third time to try to elicit a single response; otherwise, a failing performance was noted. Point-to-point interobserver agreement was conducted by an expert in ASD (and codeveloper of the task battery) on approximately 20% of the data ($n = 6$ tapes). Interobserver agreement was excellent on all control and test questions (range = 97.2%–100%; $M = 99.5\%$).

Statistical Considerations

To examine test-retest reliability, the nine tasks, consisting of 16 questions, were evaluated by the use of conservative estimates of reliability that scrutinized exact agreement and that corrected for chance agreement by a method borrowed from Mayes et al. (1996). Proportion data for each of the 16 questions were submitted to a rows-by-columns contingency table such as the one presented below:

	Time 2 Incorrect	Time 2 Correct	Marginal Proportions
Time 1 incorrect	A	B	G_1
Time 1 correct	C	D	G_2
Marginal proportions	F_1	F_2	

A, B, C, D = proportion of cases in each cell.

Test-retest reliability was calculated using the following kappa statistic (Fleiss, Cohen, & Everitt, 1969), which is an index of agreement that corrects for chance agreement:

The proportion of observed agreement (PO) = $A + D$
The proportion of chance agreement (PC) = $F_1G_1 + F_2G_2$
kappa = (PO − PC)/(1 − PC)

Further analysis of the reliability between sessions was conducted by examining separately the PO between sessions on correctly answered questions (P_{pos}) and PO between sessions on incorrectly answered questions (P_{neg}) (Cicchetti & Feinstein, 1990). The formulas and criteria adopted by Mayes et al. (1996, p. 315) follow:

$$P_{pos} = D/([F_2+G_2]/2)$$
$$P_{neg} = A/([F_1+G_1]/2)$$

Using PO, P_{neg}, P_{pos}, and kappa, a clinically useful criterion for acceptable interrater reliability is a PO, P_{neg}, and P_{pos} value of .70, with a corresponding kappa value of at least .40. Reliability is unacceptable when one or more of these indices is less than their acceptable level.

Results

Descriptive Statistics

The number and proportion of participants passing the test questions for each task at the two times and the mean for Time 1 (T1) and Time 2 (T2) are presented in Table 2. These data are based on group performance and may be construed as an index of item difficulty.

Table 2
Number and Proportion (%) of Participants Passing Each Question on the Theory-of-Mind Task Battery at Time 1 (T1), Time 2 (T2), and the Mean of T1 and T2

Item	T1 (%)	T2 (%)	M^a (%)
1a	15/18 (83)	16/18 (89)	31/36 (86)
1b	15/18 (83)	16/18 (89)	31/36 (86)
1c	16/18 (89)	16/18 (89)	32/36 (89)
1d	16/18 (89)	17/18 (94)	33/36 (92)
2	14/18 (78)	15/18 (83)	29/36 (81)
3	14/18 (78)	14/18 (78)	28/36 (78)
4a	8/18 (44)	12/18 (67)	20/36 (56)
4b	5/18 (28)	9/18 (50)	14/36 (39)
4c	8/18 (44)	11/17 (65)	19/35 (54)
5a	15/18 (83)	15/18 (83)	30/36 (83)
5b	7/18 (39)	7/18 (39)	14/36 (39)
6a	13/18 (72)	13/17 (76)	26/35 (74)
6b	9/18 (50)	11/16 (69)	20/34 (59)
7	9/17 (53)	11/18 (61)	20/35 (57)
8	7/18 (39)	9/18 (50)	16/36 (44)
9	2/18 (11)	4/18 (22)	6/36 (17)

[a]Overall difficulty rating for each question.

Inferential Statistics

Test-retest reliability by item. Table 3 organizes the various indices of agreement between responses at T1 and T2 to address the question of test-retest reliability. Values are presented for the proportion of observed agreement (PO), the proportion of chance agreement (PC), the kappa statistic, the proportion of observed negative agreement (P_{neg}), and the proportion of observed positive agreement (P_{pos}).

Ten of the questions on the ToM task battery achieved acceptable reliability, on the basis of the criteria cited above, and six of the questions did not. The questions that did not achieve acceptable reliability on the basis of the kappa analysis were 1c (emotion recognition "mad"), 1d (emotion recognition "scared"), 2 (inference of situation-based emotion), 5b (line-of-sight, visual perspective taking), 6b (inference of action based on perception), and 9 (second-order false belief task). It is immediately evident that the two easiest and the most difficult items in the task battery, on the basis of overall proportion of participants passing each question (Table 2), are included among the six items with unacceptable reliability. For Questions 1c and 1d, in which children pointed to a picture of a mad face and scared face, 89% and 92% of responses were correct, respectively. For Question 9, a second-order false belief task including a lengthy narrative and two control questions, only 17% of the responses were correct (11% at T1 and 22% at T2). On examination, it was apparent that these were the only three items with a zero or 1 in one of the agreement cells (A or D), meaning that zero or 1 out of 17 participants responded incorrectly (1c and 1d) or correctly (9) at both administrations. This reflects an extreme imbalance in the distribution of responses for these three questions.

It is recognized that problems arise for indices of agreement that correct for chance agreement (e.g., kappa) when the frequency of a target behavior is very high or very low and that the probability of chance agreement increases dramatically as this imbalance increases (McReynolds & Kearns, 1983). Feinstein and Cicchetti (1990), referring specifically to the distribution of data in the types of contingency tables used here, explained that large imbalances result in a high proportion of chance agreement (PC). Because kappa is a statistical procedure to correct for chance agreement, a high PC results in a low value of kappa in spite of a high proportion of observed agreement (PO).

Cicchetti and Feinstein (1990) offered a criterion for judging PO that appears useful in cases in which extreme

Table 3
Reliability Indices for the 16 Questions on the Theory of Mind Task Battery

Question	Cell A T1 Fail T2 Fail	Cell B T1 Fail T2 Pass	Cell C T1 Pass T2 Fail	Cell D T1 Pass T2 Pass	PO	PC	Kappa	P_{neg}	P_{pos}	A/NA Reliability
1a	2	1	0	14	0.941	0.750	0.764	0.810	0.965	A
1b	2	1	0	14	0.941	0.750	0.764	0.810	0.965	A
1c	1	1	1	14	0.882	0.792	0.433	0.500	0.933	A[a]
1d	0	2	1	14	0.824	0.837	?0.080	0.000	0.904	A[a]
2	2	2	1	12	0.824	0.671	0.465	0.572	0.889	NA
3	4	0	0	13	1.000	0.640	1.000	1.000	1.000	A
4a	6	4	0	7	0.765	0.475	0.552	0.750	0.779	A
4b	9	4	0	4	0.765	0.516	0.515	0.818	0.704	A
4c	6	3	0	7	0.813	0.485	0.637	0.800	0.825	A
5a	3	0	0	14	1.000	0.709	1.000	1.000	1.000	A
5b	7	3	4	3	0.588	0.525	0.133	0.667	0.461	NA
6a	4	1	0	10	0.938	0.589	0.849	0.893	0.957	A
6b	4	3	2	7	0.667	0.507	0.325	0.617	0.705	NA
7	5	3	1	7	0.750	0.400	0.583	0.715	0.778	A
8	9	2	0	6	0.882	0.508	0.760	0.900	0.857	A
9	13	2	1	1	0.824	0.748	0.302	0.897	0.401	A[a]

Note: A = acceptable reliability (criteria: PO, P_{neg}, P_{pos} > 0.700; kappa > 0.400); NA = not acceptable reliability; A[a] = acceptable level of agreement (PO > 0.700) according to Cicchetti & Feinstein (1990), which may be taken as an alternative indicator of reliability when kappa statistics do not capture agreement because of extreme imbalances in marginal proportions; PO = proportion of observed agreement between sessions; PC = proportion of chance agreement between sessions; kappa = index of agreement corrected for chance; P_{neg} = proportion of observed agreement on incorrectly answered questions; P_{pos} = proportion of observed agreement on correctly answered questions.

imbalances (as is often the case for the easiest and most difficult questions) result in inaccurate estimates of reliability. They suggest that 80% to 89% PO is "good" and 70% to 79% PO is "fair" agreement. Following this guideline, Questions 1c and 1d may be classified as demonstrating good agreement in the absence of failing performance (P_{neg}), and Question 9 may be characterized as demonstrating fair agreement in the absence of passing performance (P_{pos}).

The remaining three questions demonstrating unacceptable reliability according to the criteria for kappa analysis produced less extreme imbalances in their contingency tables. Therefore, the finding of unacceptable reliability for Questions 2, 5b, and 6b is considered more accurate and, thus, meaningful to the question of test-retest reliability for this ToM task battery.

Internal consistency. Internal consistency was examined using Cronbach's alpha, which is appropriate for dichotomously and nondichotomously scored items (McCauley, 2001). According to conventional guidelines, an alpha of .70 was considered "adequate," an alpha of .80 was considered "good," and an alpha of .90 was considered "excellent." Analysis revealed that internal consistency of this 16-question battery achieved $\alpha = .91$ at T1 and $\alpha = .94$ at T2, representing excellent intertask agreement. For both testing times, dropping Question 5b (which has previously been identified as unreliable) resulted in increases in alpha.

Test-retest reliability and interval length. To evaluate the effect of length of test-retest interval, comparisons were conducted between shorter (2–7 weeks) and longer intervals (8–16 weeks) in terms of change in score. An independent samples *t* test revealed no effect. Consistent with this result, a Pearson's *r* revealed no relation between change in score from T1 to T2 and the length of the interval between administrations. Thus, analyses of both differences and correlations converged to find that variation in interval did not significantly affect test-retest reliability.

Test-retest reliability and VMA. To explore whether reliability varied as a function of VMA, participants were divided into two groups according to their change in score between T1 and T2. "Consistent performers" (change of 0–1 points) and "inconsistent performers" (change of 2 or more points) were compared in terms of VMA. An independent samples *t* test revealed no effect. To explore whether consistency of performance varied as a function of diagnosis, the change in score from T1 to T2 was compared among children diagnosed with autism ($n = 8$) and PDD-NOS ($n = 7$). Because only two children in the current sample were diagnosed with

Asperger's Disorder, data for these cases were dropped from this analysis. An independent samples t test revealed no effect of diagnosis on reliability. Incidentally, the mean for the two children with Asperger's Disorder was nearly identical to the means of the other two groups. Reliability and VMA were also examined by the use of a Pearson's r, which included the full sample of participants. No relationship was found between VMA and change in score between T1 and T2. In summary, analyses of differences and correlations converged to find that variation in VMA was not associated with variation in reliability.

Discussion

Despite the popularity of ToM tasks to assess the social understanding of individuals with ASD, no study to date has examined the test-retest reliability of these tasks. The first goal of this study was to examine the test-retest reliability and internal consistency of ToM tasks representing a range of content and complexity when adapted to support the performance of children with ASD. Of the 16 items on the task battery, 10 demonstrated adequate test-retest reliability in an analysis using kappa, and three others met an alternative criterion for adequate agreement that can be used when gross imbalances in cell frequencies occur because of extreme ease or difficulty of an item.

In consideration of the three questions that did not demonstrate acceptable test-retest reliability, a question is raised about what may have contributed to the lack of reliability and what revisions might yield improvement. In Question 2, children were asked to infer an emotion on the basis of a situation. A short illustrated narrative was presented that included an explanation that Carlos doesn't like it when his sister plays with his toys because she doesn't give them back and this makes him mad. Participants were then told that today, Carlos's sister began playing with his toys. The question was posed, "How will this make Carlos feel?" and Carlos was shown making a sad, mad, happy, and scared face. Although this question did not appear to be particularly difficult, with 81% of respondents correctly choosing mad, it elicited enough change in responses (typically by selecting the "sad" response alternative) to be judged unreliable. It may be that some children were less consistent when responding to this item because attitudes conveyed by caregivers have the potential to create in children a response bias toward more socially acceptable mental states and emotions, such as a bias toward sad as opposed to mad. A competing explanation is that fluctuations in motivation, pragmatics, language, and executive function among children with ASD may interact with task performance and therefore operate unevenly.

For Question 5b, none of the reliability indices met the criteria for acceptable test-retest reliability. This was the second question used to assess visual perspective taking. Children were shown an illustration of a candle and were asked to point first to one of four images (each showing the same candle rotated successively 90 degrees) that represented what they saw when looking at the illustration. They were then asked to point to one of four images that represented what the examiner saw looking at the picture from the opposite side of the table. This question was found to be particularly confusing, and on closer examination, two competing interpretations may be possible. One response (the incorrect one) is plausible if the task is interpreted as "what visual or retinal image corresponds to the comparison illustration that the other person sees right now?" By contrast, the correct response requires the respondent to interpret the task as "what image would correspond to the comparison illustration that I see if I were to trade places with the other person and see what she sees right now?" Therefore, the poor reliability of this question may reflect ambiguous wording. Future studies might include a three-dimensional object in place of an illustration as the stimulus for assessment of line-of-sight perspective taking (e.g., Piaget's classic three pyramids task). Another option might be to use a barrier to obstruct the views of the different observers (e.g., Hadwin et al., 1996). In this way, test questions could be posed in such a way that there is no competing interpretation, which in turn should contribute to reliable responding.

Question 6b is the final item that did not meet the criteria for adequate test-retest reliability. The narrative for this question stated, "This is Patty. This morning Patty saw her glasses on the table. Now she wants her glasses." Question 6a, measuring the ability to infer belief-based perception, asked, "Where does Patty think her glasses are?" Of the responses to this question, 74% were correct, and the item achieved acceptable reliability. Question 6b was designed to measure the ability to infer action-based perception by asking, "Where will Patty look for her glasses?" Only 59% of the responses to this question were correct, and the item did not achieve acceptable reliability. Question 6a used a mental state term with an embedded complement, which is a linguistic structure considered more challenging for children with ASD compared with the structure of 6b (Tager-Flusberg, 2000). For this reason, it might be expected that 6b would be easier and more reliable than 6a. What then, would explain the relative difficulty and inadequate

reliability of 6b? It may be that an unusual pragmatic situation occurs when one is asked two different questions in succession that require the same response. This may cause confusion because the respondents may have inferred that the two answers should be different. If this interpretation is correct, unacceptable reliability for this item can be, in part, attributed to the children's pragmatic knowledge as opposed to pragmatic deficits.

It is noteworthy that, although results were mixed, Mayes et al. (1996) also reported unacceptable reliability for some of their "looking" items that immediately followed "thinking" items. As in this study, those questions were asked in succession and required the same response for success. This issue is of critical importance. Repeated questioning is a common component of ToM tasks and is thought to encourage answer switching, which might be especially problematic for larger task batteries (Wellman et al., 2001). In future research, the order of these two questions could be counterbalanced to examine how order affects reliability. Alternatively, the ability to infer belief-based perception and action-based perception should be assessed using two separate scenarios.

Analysis of internal consistency indicated very good intertask agreement. Previous studies of ToM tasks administered to typically developing children found internal consistency within (Hughes et al., 2000) and across studies (Wellman et al., 2001), and some have concluded that ToM is a "unitary construct" (Hughes et al., 2000, p. 486). However, the evidence that the various tasks employed to assess ToM competence converge in a meaningful way can also support the notion of a multidimensional set of related abilities, which seems to be a more accurate interpretation if one conceives of ToM as a broad construct that represents a large set of ToM knowledge areas. Future research should employ vastly larger samples to allow for factor analysis of large ToM task batteries in order to reveal the nature of the relationships between items and how groups of items may relate to different aspects of a multidimensional ToM construct.

The second goal of this study was to examine whether reliability differed between relatively short and long test-retest intervals. Analyses revealed no effect of interval on change in score between T1 and T2. It is noteworthy that, when scores did change, the change was modest and usually (but not always) in a positive direction. However, these changes were not significant, suggesting that maturation and test-retest effects did not play a substantial role. These results are consistent with Mayes et al. (1996), who employed a much shorter test-retest interval when assessing the performance of typically developing children.

With regard to the third goal of this study, it was found that the reliability of performance did not vary as a function of VMA, and this is consistent with a previous study of typically developing children (Hughes et al., 2000). However, a relationship between VMA and consistent (or reliable) performance on ToM tasks was evident in a study of children with learning disabilities (Charman & Campbell, 1997). These conflicting results underscore the need for establishing the psychometric properties of measures across populations. Because reliability did not vary as a function of VMA, it is not surprising that there was no effect found of diagnosis on reliability. Although investigations should be extended to larger samples, this finding suggests that the reliable items on the current ToM task battery are appropriate to administer to individuals across the autism spectrum.

Crucially, researchers question the validity of false belief tasks for individuals with both ASD and limited verbal skills because "they lack the cognitive and verbal skills necessary to answer the control questions, success on which is usually an inclusion criterion" (Happe, 1995, pp. 845–847). These concerns would certainly apply to at least three children in the current sample, who were identified as functionally nonverbal and who evidenced scores on a standardized test of receptive vocabulary corresponding to a VMA of less than 3 years. Of these three children, one attended to the task but could not demonstrate understanding of the most basic tasks (e.g., emotion recognition) or the control and test questions of more advanced tasks. Another child demonstrated understanding of basic tasks but failed the control and test questions of more advanced tasks. The third child demonstrated understanding of basic tasks and understanding of control questions for advanced tasks but failed test questions for the advanced tasks. These reliable and variable patterns of performance underscore variation in the performance of children with the most limited verbal abilities and are important from research and clinical perspectives.

These findings corroborate those of Happe (1995), who cautioned against the use of ToM tasks for children with very limited verbal abilities who cannot demonstrate understanding of control questions. In this situation, such tasks probably tap language ability, pragmatic understanding of the testing situation, motivation, or other constructs irrelevant to ToM competence. But when the control questions are passed (reliably) and test questions are not or when basic emotion recognition tasks are passed but advanced tasks are not, the performance patterns remind us that ToM knowledge should not be assumed absent among nonverbal children with ASD, who have traditionally been excluded from examinations of ToM competence. Variable patterns of performance such as these may reveal the aspects of ToM knowledge (and receptive language) that represent relative strengths

and weaknesses for the individual with ASD. As such, ToM tasks can guide clinical impressions of the social understandings for which an individual may have basic knowledge but exhibits a performance deficit. Such tasks may also facilitate clinical thinking regarding ways to promote and assess ToM competence over time.

Of course, some limitations of this study warrant mention. Chief among them was the use of a relatively small sample size. A larger sample size would enhance confidence in the stability of the findings and would have allowed for statistical procedures (e.g., Guttman and Rasch analyses) to examine whether tasks reflect a sequence of understanding evident in children's developing ToM as this has potential implications from scalability, administration and scoring, and theoretical perspectives (Peterson, Wellman, & Liu, 2005; Wellman & Liu, 2004). In addition, although the current task battery was specifically designed to assess a range of ToM content and complexity, it is by no means considered exhaustive. A more comprehensive battery incorporating tasks to assess other areas of social understanding known to be impaired in individuals with ASD (e.g., understanding of speech acts and mental–physical and appearance–reality distinctions) would enhance the coverage of content relevant to ToM.

Despite these limitations and given the lack of previous research in this area, this study offers preliminary empirical support for the notion that a range of ToM tasks can be administered reliably when they are administered in ways that are sensitive to the challenges faced by children with ASD. The ToM task battery described here is a promising tool that can be administered to children across the autism spectrum. Keeping in mind that caution and care must be exercised in interpreting the pattern of performance on control and test questions, the tool also has the potential to inform clinical decision making for individuals who present with the most limited verbal capacities. This study also underscores the need for further investigation of the psychometric properties of the most common tasks used to assess the social understanding of individuals with ASD. The generally good reliability of items in this study should not be taken to support the reliability of ToM tasks in general when applied to children with ASD. Recall, the nature of the materials, stimuli, and task administration procedures in this study were designed so that they were sensitive to the ways children with ASD process information and perform best. It is worthwhile now to explore variations in administration procedures to establish the extent to which the reliability of ToM tasks may be generalized across a broader range of tasks and data collection procedures.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Baron-Cohen, S. (1991). Do people with autism understand what causes emotion? *Child Development, 62*, 385–395.

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

Beukelman, D., & Mirenda, P. (1992). *Augmentative and alternative communication: Management of severe communication disorders in children and adults*. Baltimore: Paul H. Brookes.

Bishop, D. (1989). *The Test for Reception of Grammar.* London: Medical Research Group.

Charman, T., & Campbell, A. (1997). Reliability of theory of mind task performance by individuals with a learning disability. *Journal of Child Psychology & Psychiatry, 38*, 725–730.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*, 551–558.

Dettmer, S., Simpson, R. L., Myles, B. S., & Ganz, J. B. (2000). The use of visual supports to facilitate transitions of students with autism. *Focus on Autism and Other Developmental Disabilities, 15*, 163–169.

Dunn, L. M., & Dunn, L. M. (1997). *Examiner's manual for the Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543–549.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further knowledge about the Level 1–Level 2 distinction. *Developmental Psychology, 17*, 99–103.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323–327.

Grandin, T. (1995). The learning styles of people with autism: An autobiography. In K. A. Quill (Ed.), *Teaching children with autism: Strategies to enhance communication and socialization* (pp. 33-52). Albany, NY: Delamar.

Hadwin, J., Baron-Cohen, S., Howlin, P., & Hill, K. (1996). Can we teach children with autism to understand emotions, belief, or pretence? *Development and Psychopathology, 8*, 345–365.

Happe, F. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*, 843–855.

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

Johnston, S., Nelson, C., Evans, J., & Palazolo, K. (2003). The use of visual supports in teaching young children with Autism Spectrum Disorder to initiate interactions. *Augmentative and Alternative Communication, 19*, 86–103.

Kaplan, R. M., & Saccuzzo, D. P. (1989). *Psychological testing: Principles, applications, and issues.* Pacific Grove, CA: Brooks/Cole.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review, 94*, 412–426.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (1999). *Autism Diagnostic Observation Schedule-Generic (ADOS-G).* Los Angeles: Western Psychological Services.

Mayes, L., Klin, A., Tercyak, K. P., Cicchetti, D. V., & Cohen, D. J. (1996). Test-retest reliability for false-belief tasks. *Journal of Child Psychology and Psychiatry, 37,* 313–319.

McCauley, R. (2001). *Assessment of language disorders in children.* Mahwah, NJ: Lawrence Erlbaum.

McReynolds, L. V., & Kearns, K. P. (1983). *Single-subject experimental designs in communicative disorders.* Baltimore: University Park Press.

Mitchell, P., Saltmarsh, R., & Russell, H. (1997). Overly literal interpretations of speech in autism: Understanding that messages arise from minds. *Journal of Child Psychology and Psychiatry & Allied Disciplines, 38,* 658–691.

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, 67–80.

Nunnally, J. C., & Berstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development, 60,* 689–700.

Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development, 76,* 502–517.

Prior, M., Dahlstrom, B., & Squires, T. L. (1990). Autistic children's knowledge of thinking and feeling in other people. *Journal of Child Psychology & Psychiatry, 31,* 587–601.

Siegal, M., & Beattie, K. (1991). Where to look first for children's knowledge of false beliefs. *Cognition, 38,* 1–12.

Silliman, E. R., Diehl, S. F., Bahr, R. H., Hnath-Chisolm, T., Zenko, C. B., & Friedman, S. A. (2003). A new look at performance on theory-of-mind tasks by adolescents with Autism Spectrum Disorder. *Language, Speech, and Hearing Services in Schools, 34,* 236–252.

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology, 30*(3), 395–402.

Tager-Flusberg, H. (2000). Language and understanding minds: Connections in autism. In S. Baron-Cohen, H. Tager-Flusberg, & D. H. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience* (2nd ed., pp. 124–149). Oxford, UK: Oxford University Press.

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

Wechsler, D. (1990). *Wechsler Preschool and Primary Scale of Intelligence–Revised.* London: Psychological Corporation, Harcourt Brace.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false beliefs. *Child Development, 72,* 655–684.

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

Williams, K. T., & Wang, J. (1997). *Technical references to the Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

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

Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin, 124,* 283–307.

Tiffany L. Hutchins, PhD, is a lecturer of communication sciences at the University of Vermont. Her current interests include autism, test development, and the relation between maternal beliefs and parenting practices.

Patricia A. Prelock, PhD, CCC-SLP, is a full professor and chair of the Department of Communication Sciences at the University of Vermont. Her research interests focus on social communication assessment and intervention in children with ASD.

Wendy Chace, MS, CCC-SLP, works for a home health agency in rural Vermont, serving clients with a variety of communication impairments and focusing on early intervention (birth to 3 years).