CHILD DEVELOPMENT

Child Development, March/April 2012, Volume 83, Number 2, Pages 469–485

# The Mind Behind the Message: Advancing Theory-of-Mind Scales for Typically Developing Children, and Those With Deafness, Autism, or Asperger Syndrome

Candida C. Peterson
*The School of Psychology at University of Queensland*

Henry M. Wellman
*University of Michigan*

Virginia Slaughter
*The School of Psychology at University of Queensland*

Children aged 3–12 years ($n$ = 184) with typical development, deafness, autism, or Asperger syndrome took a series of theory-of-mind (ToM) tasks to confirm and extend previous developmental scaling evidence. A new sarcasm task, in the format of H. M. Wellman and D. Liu's (2004) 5-step ToM Scale, added a statistically reliable 6th step to the scale for all diagnostic groups. A key previous finding, divergence in task sequencing for children with autism, was confirmed. Comparisons among diagnostic groups, controlling age, and language ability, showed that typical developers mastered the 6 ToM steps ahead of each of the 3 disabled groups, with implications for ToM theories. The final (sarcasm) task challenged even nondisabled 9-year-olds, demonstrating the new scale's sensitivity to post-preschool ToM growth.

Theory of mind (ToM)—the explicit understanding of how human behavior is governed by mental states of belief, intention, memory, and desire—develops rapidly for most children during the preschool years. Most 3-year-olds fail to demonstrate explicit ToM-based awareness of representational mental states, as assessed prototypically using false belief (FB) tests that require predictions or explanations about the actions or thoughts of protagonists with beliefs that are out of line with reality. Yet, by 4 or 5 years, typically developing children pass so consistently as to suggest that ''understanding belief and, relatedly, understanding of mind, exhibit genuine conceptual change during the preschool period'' (Wellman, Cross, & Watson, 2001, p. 655).

The developmental picture is complicated and less understood for children with developmental delays, such as autism (for reviews, see Baron-Cohen, 2000; Happé, 1995; Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998), deaf children of hearing parents (though not, interestingly their deaf peers whose parents are deaf signers; Peterson, 2009; Siegal & Peterson, 2008), as well as blind children (Siegal & Peterson, 2008), and those with developmental challenges such as severe maltreatment (Sharp & Fonagy, 2008). One limitation of the research with delayed and typically developing children alike is its overwhelming focus on FB. A genuine understanding of others' minds cannot be equated with performance on standard inferential FB tests alone; ToM understanding evidences a progression of insights that unfold over development (Pons, Lawson, Harris, & deRosnay, 2003; Wellman & Liu, 2004). Comparisons across groups of children from different backgrounds, in particular, require more comprehensive developmental data.

In response to such concerns, Wellman and Liu (2004) devised a developmental scale of ToM that assesses multiple milestones in the growth of social cognition. The scale charts five sequential steps in explicit understanding of minds, using carefully constructed tasks that match one another closely in linguistic and procedural demands and their

The authors thank the children who participated and their parents and teachers. This research was partly supported by an Australian Research Council Discovery grant to Peterson, Slaughter, and Wellman and a U.S. National Institute of Health grant to Wellman.

Correspondence concerning this article should be addressed to Candida C. Peterson, School of Psychology, University of Queensland, Brisbane, Queensland 4072, Australia. Electronic mail may be sent to candi@psy.uq.edu.au.

0009-3920/2012/8302-0009
DOI: 10.1111/j.1467-8624.2011.01728.x

overall format and scoring. In brief, the specific tasks comprising the scale, are (a) diverse desires (DD; different people want different things), (b) diverse beliefs (DB; different people have contrasting, potentially true, beliefs about the same thing), (c) knowledge access (KA; not seeing leads to not knowing), (d) false belief (FB; standard misleading container task), and (e) hidden emotion (HE; people can feel a different emotion from the one they display). Research with several hundred U.S. preschoolers confirms a reliable five-step developmental progression that, with important variation, generalizes to children growing up in different countries and language communities (e.g., Kristen, Thoermer, Hofer, Aschersleben, & Sodian, 2006; Wellman, Fang, Liu, Zhu, & Liu, 2006), and to ToM-delayed children with autism or deafness (Peterson & Wellman, 2009; Peterson, Wellman, & Liu, 2005; Remmel & Peters, 2009).

Given this prior research and the scale's potential to provide an extended metric of development, several questions arise. Most focally, can the scale be extended further to older children who have already mastered FB? Would an extended scale likewise apply reliably to children with deafness or autism? If so, this new metric could assist in further ToM research with older typically and atypically developing groups—for example, in helping to explain why, even after mastering FB, many preadolescent children with autism continue to suffer severe problems with social and peer relations (Dissanayake & Macintosh, 2003; Peterson, Slaughter, & Paynter, 2007). Less focal to these primary aims (but still of theoretical and practical interest) would be to use such a scale to explore developmental progressions for children with Asperger syndrome (AS), a diagnostic group never previously included in ToM scaling research.

### *Later ToM Achievements*

Potentially, any number of tasks could be harder, thus later developing, than those in the current preschool scale. Indeed, any number of additional tasks could measure preschool ToM as well. In keeping with the original scale's use of a small set of representative but strictly scalable tasks, we sought a single additional task that could extend the scale to older (i.e., school-aged) children. Conceptually, we sought a task related to everyday social-cognitive demands but also one in keeping with the current scale's overall emphasis on understanding differences between mental states across people (e.g., self vs. other) and differences between mental states and reality (e.g., knowledge vs. ignorance or felt vs. expressed emotions).

One promising everyday social problem that may demand a more sophisticated level of mind-reading skill than false belief or hidden-emotion understanding is the appreciation of a speaker's (or writer's) communicative intent in situations like sarcasm, humor, and irony. Such nonliteral communicative situations conceptually require that the listener ''apprehend the mind that lies behind the message'' (Rajendran, Mitchell, & Rickards, 2005, p. 434). Happé (1994) included both irony and sarcasm in the items composing her ''advanced'' strange stories ToM test, while Filippova and Astington (2008) similarly noted that ''understanding higher order representations of belief, intention, and emotion is required . . . to comprehend indirect speech acts'' (p. 126).

Indeed, there is some empirical evidence that, for typically developing children, the comprehension of irony or sarcasm both follows and partially depends upon earlier ToM mastery. Filippova and Astington (2008) compared children aged 5, 7, and 9 and an adult control group on eight items all requiring explanations for story characters' ironic statements (e.g., saying ''You're a GREAT scorer'' when someone misses an easy football kick). Performance was fairly consistent across eight different story items but clear age differences emerged. The 5-year-olds often scored no better than chance, taking the ironic remark at its literal face value, and believing the speaker meant it as such. By age 7, many children recognized that the utterance was not literally true but failed to appreciate the speaker's motivational attitude. Even at age 9 only 25% of children understood the irony fully enough to recognize that the speaker's pragmatic purpose was ''teasing, joking or being sarcastic'' (p. 138). Standard ToM tests (Banerjee, 2000; Perner & Wimmer, 1985) were given to assess the children's understanding of second-order FB and related concepts. Scores on these correlated with irony/sarcasm scores at the univariate level and continued to make a modest (though not always significant) contribution once the influences of age, receptive vocabulary, digit span, and prosody detection were taken into account.

For typically developing children, then, there are good empirical grounds for believing that understanding nonliteral, ironic, or sarcastic messages is a more advanced aspect of social cognition. However, despite some evidence for connections among discrete pairs of tasks, no previous research has mapped sarcasm's place relative to a systematic progression of ToM understandings.

### *Delayed Groups With Deafness or Autism*

Gregory, Bishop, and Sheldon's (1995) interviews with the hearing parents of deaf young adults who had grown up in hearing households revealed persistent adult difficulties with nonliteral language and sarcastic humor even among those who were functioning quite successfully both as mature communicators (in speech or sign) and in everyday life within their communities. One hearing mother reported that her 19-year-old daughter, a British Sign Language (BSL) user, ''doesn't know the meaning of a joke; if you say something, it's serious. She can't see a double meaning . . . as far as language goes, you can't play around with it'' (p. 33). In general, verbal humor and sarcasm posed problems for 56% of this sample of severely or profoundly deaf young adults, with no distinction between signers (of BSL or Signed English) and oral-language users. Unfortunately, no measures of ToM understanding were included in Gregory et al.'s study. At the same time, however, independent longitudinal evidence (e.g., Peterson, 2009; Wellman, Fang, & Peterson, 2011) suggests that many deaf children and adolescents of hearing parents do, very belatedly, manage to master simpler ToM concepts. In particular, by the ages of 10–12 years, many are finally able to pass false belief and hidden-emotion items from the Wellman and Liu scale (Wellman et al., 2011). Together with Gregory et al.'s findings, this suggests that understanding ironic, sarcastic expressions might be a still later and problematic achievement and one with significant real-life importance for social interaction.

In parallel, there is evidence from children with autism to show that problems understanding sarcasm and irony persist even amongst the older, very high-functioning minority with exceptionally advanced language skills who eventually master basic ToM concepts of FB. For example, Happé (1994) compared subgroups of children and adults with autism who had passed ($n = 12$) versus failed ($n = 6$) standard FB tasks in terms of their ability to explain why story characters made a series of literally untrue statements, including comments that were sarcastic or ironic (e.g., saying ''That's really polite!'' to someone who has been rude). Both subgroups were outperformed by typically developing adult controls (who scored nearly perfectly). The autistic FB passers in this small sample were significantly more likely to pass sarcasm and irony items than others with autism who failed FB.

Similarly, Jolliffe and Baron-Cohen (1999) compared groups of participants with high-functioning autism, AS (see next section), or no disability on a set of 18 nonliteral language stories that included the irony and sarcasm stories described above. Despite their perfect scores on a standard first-order FB test, the groups with autism and AS were significantly poorer at the nonliteral language items than the controls. Thus, even for these unusually advanced individuals on the autism spectrum, comprehension of nonliteral statements seemed to require development beyond FB. A limitation of the study, however, was failure to separate nonliteral language forms like metaphor, that may require literacy (a non-ToM skill that individuals on the autism spectrum may also be slow to acquire) from the sarcastic and ironic remarks that often arise in social conversation.

Our primary aim was to include sarcastic irony to attempt to extend the preschool ToM Scale one further step to create an expanded scale usable with older children. To best assess the nature and usefulness of this proposed scale expansion, we tested it primarily with typically developing children, deaf children of hearing parents, and children with autism—three contrasting groups for which the original five-item scale has proved reliable and informative. We also explored possible use of the scale with children with AS.

### *Asperger Syndrome*

AS is a disorder on the autism spectrum that shares with classic autism the problems of impaired imagination/cognitive rigidity and poor social communication/reciprocity, while lacking the severe language delays that complete autism's clinically diagnostic triad (American Psychiatric Association [APA], 2000). Despite this, research on whether or not ToM development is delayed for children with AS is inconclusive. Few direct comparisons between AS and autism even on FB tasks have been published and their findings are mixed. Dahlgren and Trillingsgaard (1996) discovered both these groups scored equally highly (85%–90%) on a first-order FB test and, indeed, equaled mental age-matched controls. However, on a second-order FB task demanding a higher level of recursive processing (e.g., ''He thinks that she thinks that X''; Perner & Wimmer, 1985), the typical controls outperformed those with AS and autism, who continued to score equally (60% success). Ozonoff, Rogers, and Pennington (1991) similarly found no significant difference between groups with AS versus autism on first-order FB once language differences (vocabulary size) were statistically controlled. By contrast, two more recent studies using batteries of FB tasks (Dissanayake & Macintosh, 2003; Paynter & Peterson, 2010) found better first-order false belief performance by those with AS than autism, even after controlling for verbal ability. Additionally, and contradictorily, Paynter and Peterson (2010) found their AS sample performed equally to age- and language-matched typically developing children, while Dissanayake and Macintosh (2003) found younger typical developers mastered FB ahead of older children with AS.

Beyond this limited evidence for FB tasks alone, the overall question of the extent to which individuals with AS are delayed in ToM understanding more broadly remains almost unstudied, both in relation to typically and atypically developing peers. Indeed, it may not be possible to answer questions about ToM delay in AS using FB tests alone. Owing to the absence of language delay from its diagnostic criteria (APA, 2000), AS is often not discovered until late primary school, at ages when FB is no longer optimally sensitive to individual differences in social understanding, at least for typically developing children who often score at ceiling. Consequently, for children with AS in particular, we need more developmental data on more advanced tasks in order to capture an extended progression of ToM achievements. More generally, for those with autism and deafness, as well as AS, we know too little of their extended trajectories of ToM development, amidst delays, both up to and beyond the traditional litmus criterion of FB understanding.

### *Current Study*

In short, more sensitive and encompassing developmental comparisons of ToM between typically developing children and those with deafness and autism are needed, and especially for ages and developmental levels beyond the social-cognitive insights achieved by young preschoolers. Thus, in the current research, we adopt a scaling approach but extend the prior ToM Scale for use with older children. To do so, we focus on children's understanding of a particular type of nonliteral message that is a familiar part of children's everyday family and peer conversation, an understanding of sarcastic irony. Among typical developers, we concentrate not only on preschoolers but also on those aged 6–12, a group not previously tested on the Wellman and Liu (2004) scale. We also study fresh samples from populations assessed on the scale previously and shown to encounter theoretically intriguing delays, namely, children with autism and deaf children of hearing parents. These groups are of special theoretical interest in light of Peterson et al.'s (2005) findings (a) that deaf children from hearing families were substantially delayed behind hearing children in scale progress, although their deaf peers and classmates who were native signers (having grown up with signing deaf parents) progressed through the ToM Scale as rapidly as typical developers, and (b) that children with autism progressed through the scale steps with the same overall degree of delay as deaf children of hearing parents but in a different order. Given its potential relevance to theoretical accounts of how ToM is acquired (e.g., Harris, 2006; Siegal & Peterson, 2008; Tager-Flusberg, 2003), it is important to examine the possibility of ToM delays and possible divergences and to do so across a broader scale of ToM growth.

To summarize our multiple aims, first, we explore the feasibility of extending the earlier five-step ToM Scale developed for typically developing preschoolers to span six reliably sequential steps. Second, we compare typically developing children's scale progress to that of atypical developers after controlling for influences of age and language ability. Third, we examine five- and six-step Guttman scale assessments of ToM for two diagnostic groups of special theoretical interest—deaf children of hearing parents and children with autism—to see if six sequential steps are reliably scalable for these groups and also whether a unique five-step sequence previously noted for children with autism is replicable. We also, for the first time, compare extended ToM development by these two theoretically significant groups to their age-matched (as well as preschool) typically developing peers. Finally, although less focally, we examine ToM Scale development in children with AS (a novel diagnostic group not previously tested in scaling studies) to explore how their scale progressions and task patterns compare with those for children with autism, on one hand, and those for typical developers, on the other.

## Method

### *Overview*

We administered the Wellman and Liu (2004) five-item ToM Scale, together with a test of understanding of sarcastic irony and a standardized language ability test. We selected this sarcasm/irony task because it could be adjusted to a format closely parallel to that used in the five items of the preschool ToM Scale. In keeping with that scale, we focused on a single task to represent children's understanding of nonliteral communications (just as the prior scale uses a single task to test understanding, e.g., of diversity of desires, or of FBs). This feature of the prior scale, and of the current extension, means that it can be conducted in a single short (15 min) session. Thus, the scale lends itself for use with delayed groups who may have difficulties with more extended testing sessions, limited availability for research participation, or may require a broad range of assessments over and above social cognition.

### *Participants*

Four groups totaling 184 Australian children participated; aged 3–13 years. Group 1 had 31 late-signing deaf children (mean age = 9.62 years, range = 6–12; 18 boys, 13 girls), Group 2 had 44 children with autism (*M* = 9.02, range = 5–12; 37 boys, 7 girls), Group 3 had 41 children with AS (*M* = 9.52, range = 5–12; 30 boys, 11 girls), and Group 4 included 68 typically developing children subdivided into three age groups. Group 4A, the oldest, had 29 children (*M* = 8.77 years, range = 7.5–11.5; 17 boys, 12 girls). This was the primary (age matched) control for statistical comparisons of scale progress across diagnostic groups. There were no significant age differences among Groups 1, 2, 3, and 4A, $F(3, 141) = 1.52, p > .20$.

To examine a wide range of ages and ToM abilities—important for optimal Guttman scaling and for more precise mapping of ToM delays relative to normative patterns—we included younger typically developing children. Group 4B had 16 primary schoolers aged 6–7.5 years (*M* = 6.96; 8 boys, 8 girls) and Group 4C contained 23 preschoolers aged 3–5 years (*M* = 4.75; 8 boys, 15 girls).

To insure the precise diagnostic identification of the children with autism, and the differentiation between AS and autism groups, our sample was stringently selected using the accepted ''gold standard'' (Szatmari et al., 2009) of expert clinical judgment according to *DSM–IV* (APA, 2000) criteria with the added precaution that these *DSM–IV*-based differential diagnoses were conferred *and* verified by trained clinicians (including at least one qualified pediatrician or psychiatrist) working independently of the authors and blind to the outcome measures of our research. Children who received a mixed diagnosis or one that was not precisely AS or autism (e.g., ''autism spectrum disorder'' or ''PDD-NOS''), were not included. Furthermore, we excluded any deaf or typically developing child who, according to the teacher had been diagnosed or suspected of having an autism spectrum disorder or other disability (apart from hearing loss).

Children were recruited from local schools with predominantly middle-class socioeconomic catchments on the basis of their parents' written consent. No children with other diagnosed conditions (e.g., blindness, intellectual disability) were included. The deaf children all attended specialist units where a signed language (Signed English or Auslan) was the primary communication medium. Those with autism and AS attended specialist units exclusive to autism spectrum disorders. Apart from the deaf, all children and their families spoke English as their sole or primary language.

### *Tasks and Scoring*

*ToM Scale tasks*. The five-item ToM Scale (Wellman & Liu 2004)—including (a) DD, (b) DB, (c) KA, (d) FB, and (e) HE—was given precisely in the format described by Peterson et al. (2005; see See p. 517 of their article for the exact wording of all questions and instructions, together with stimuli, procedure and scoring). Conceptually, all the tasks asked about a focal contrast between a protagonist's inner psychological state (e.g., ignorance, FB, felt emotion) and either reality (e.g., true contents, overt facial expression) or the mental state of another protagonist (e.g., own vs. other's belief). All tasks were presented with the aid of drawings and/or toy figurines and all had a similar format involving a test question and at least one comprehension control question. As in Peterson et al. (2005), we required correct responses to all control as well as test questions to pass any given task. With one exception (apart from local substitutions like ''biscuit'' for ''cookie''), the tasks were identical to Wellman and Liu (2004).

The exception was the HE task where we followed Peterson et al. (2005) in using an additional explanation question and, consequently, using an alternative to the original scoring procedure (see Peterson et al., 2005, p. 517, for exact details). Stimulus pictures (e.g., sad, okay, happy face cartoons) and procedures were identical to Wellman and Liu (2004) but we added an extra comprehension control question: ''Why did he try to look [sad, okay, happy]?'' to follow-up test question responses. This verified children's understanding of the stimulus story, and disambiguated a possible confound in the original procedure. Originally, a child who had selected ''sad'' as the boy's true emotion could pass the test question simply by pointing at the ''happy'' or ''neutral'' cartoon on the three-face scale. However, pilot work with a separate sample of children and adults showed that the supposedly neutral (middle) picture was often spontaneously labeled as ''angry,'' so pointing at this face could (incorrectly) indicate negative rather than neutral emotion. Addition of the ''why'' follow-up disambiguated this. Correct responses required that any child who had merely pointed should either indicate a social motive (e.g., ''so they don't tease more,'' ''to make them stop'') or should allude to the hiding of feelings (e.g., ''so no one knows''). A reliability coder, blind to respondents' ages, group membership and performance on other tasks, independently scored the responses of 60 children who had been randomly selected from each diagnostic group. Agreement with the primary coder was 95%.

A method that includes explanations had the additional benefit of making the format for HE very comparable to that used in the new task testing understanding of the social use of nonliteral language in sarcasm. This task was modeled closely on one of the sarcasm items in Happé (1995). Rajendran et al. (2005) used a similar story which they deemed to measure both irony and sarcasm. In our version (SARC), adapted for ease of translation into sign and to insure suitable comparability in style and format to the other five ToM Scale tasks, children were told:

> The girl and boy are going on a picnic. It is the boy's idea. He says it will be a lovely sunny day. But when they get the food out, big storm clouds come. It rains and the food gets all wet. The girl says: ''It's a lovely day for a picnic.''

A colored line drawing showed the back of a girl's and a boy's head, raindrops, and a wet cake and other food on a picnic rug. The tester read the story aloud without any special intonation or emphasis. There was a preliminary question, taken from Happé (1994): ''Is it true, what the girl said?'' and a test question: ''Why did the girl say 'it's a lovely day for a picnic'?'' plus a comprehension control question new to this study: ''Was the girl happy about the rain?'' Thus the task format very closely parallels the HE task both pictorially and in its types and formats of questions: both have a control question asking about true emotion and a similar ''why'' test question.

As for the five other tasks, children had to pass the control question in order to pass SARC. Control-question failures were actually rare. Only 9 of 184 children (5%) failed the SARC control by agreeing the girl was happy about the rain or by saying ''don't know,'' and none of these children would have passed the task, even if the control were ignored.

Our criteria for passing SARC were closely modeled on Filippova and Astington's (2008) scoring of a similar ''why'' question for their sarcastic irony story. Correct ''why'' answers either mentioned sarcasm explicitly or else alluded in some other way (e.g., ''joking,'' ''doesn't mean it'') to a contrast between the literal meaning of the words ''lovely day'' and the speaker's intended meaning. This scoring conforms with widely accepted definitions of irony or sarcasm as intending to mean the ''opposite of the literal meaning of a sentence'' (Rajendran et al., 2005, p. 434). Sample responses scored as correct versus wrong are shown in the Appendix. An independent reliability coder, blind to respondents' age, gender, group membership, and the hypotheses of the study, coded a random selection of 90 responses, representing all diagnostic groups. Agreement with the primary coder was 97%.

*ToM-6 and ToM-5 summary scores*. For each of the six items children earned a score of 1 for a pass, otherwise a 0. Thus, total scores for the original five-item ToM Scale (ToM-5) ranged from 0 to 5, and for a new six-item ToM Scale (ToM-6), they ranged from 0 to 6.

*Language ability*. To assess linguistic maturity, we used the 22-item syntax subscale of the Clinical Evaluation of Language Fundamentals test (CELF–Preschool; Wiig, Secord, & Semel, 1992). This test (a) has been used effectively in earlier ToM research with typically developing children (e.g., Ruffman, Slade, & Crowe, 2002) and (b) is uniquely suitable for validly assessing linguistic maturity among signing deaf Auslan- or Signed- English users in Group 1. These deaf children cannot validly be assessed with other commonly used receptive vocabulary measures—like the Peabody Picture Vocabulary Test—for numerous reasons, including the large proportion of iconic signs (e.g., ''knee'' = pointing at the knee) in Australian signed languages. Via a picture-pointing response mode, the CELF–Preschool syntax subscale assesses a broad range of developmentally sequenced lexical, morphological, and syntactic concepts (including verb tense, relative clauses and embedded constructions). Appropriately, several of these items were found to challenge even the oldest typical developers in our sample. Thus, this CELF subscale provided an adequate range of raw score variability (4–22) not only for the sample as a whole but also within each of the six separate subgroups. (CELF scores were available for all 85 of the children with AS or autism, for 30 or the 31 deaf children [97%], and for 56 of the typical developers including, importantly, 27 [90%] in the focal age-matched subgroup). We used raw scores as the dependent measure (as advocated by the test manual for un-normed groups like ours).

### *Procedure*

Each child was tested individually by one of three highly experienced experimenters in two sessions, one for the language test and one for the ToM measures. For these latter, six different task orders were used, randomly assigned across the sample. To maximize motivation and involvement, all orders began with one of the three tasks that previous studies of children from the same populations have established are easiest, namely, DD, DB, or KA.

Owing to their preference to communicate in sign rather than speech, for deaf children the main experimenter was assisted by one of four sign language interpreters, all with professional-level qualifications in the particular sign language used in each deaf child's classroom (Signed Australian English [84%] or Auslan [16%]) as well as with each child's own particular sign language preferences. The interpreter, seated beside the experimenter in full view of the participant, translated the experimenter's speech into the child's preferred mode of sign language, using an interpretation style familiar to these children in their everyday school routines. Interpreters paused when critical pictures or figurine actions were introduced and, before continuing, both adults monitored that the child's gaze was directed at the stimuli or interpreter, as appropriate. They independently recorded pointing responses and subsequent matching of their records revealed complete agreement. The interpreter also supplied an ongoing oral translation of all the child's signed communication, which was recorded by the experimenter on data sheets. We developed tasks, test questions, and appropriate translations of the CELF items in close consultation with native speakers of Auslan and Signed Australian English to insure appropriate translation into these languages. Our interpreters were well practiced both in signed translation and in the importance of adhering exactly to the script, with the experimenter monitoring the latter.

## Results

### *Overview*

We begin by examining the data for typically developing children in order to ascertain if the SARC task is a significantly more advanced and difficult task than others on the scale, if it is passed by enough typically developing children to insure it is a genuine developmental progression, and whether it provides an appropriate and statistically reliable extension to the ToM Scale. Furthermore, given that no previous study using the five-step preschool scale has included a sample of typically developing children as old as the present one, we examined whether the earlier established progression, together with the new step, would prove replicable for this more extended sample ranging up to 11.5 years and we examined whether these children's overall ToM Scale growth is reliably predicted by age and/or language ability.

After establishing scale properties for the typically developing group, we then compared the atypically developing groups with them in scale performance. Using hierarchical regression, and after controlling statistically for variations in language ability, we assessed the independent contributions of age (a marker of progressive conceptual development) and developmental condition (e.g., autism) to children's overall pace of progress through the six-step scale. This initial information allowed us to test the Guttman scale conformity of predicted five-step and six-step developmental sequences for our two focal atypically developing groups—children with autism and deaf children of hearing parents—addressing the research questions outlined earlier, including (a) how well deaf children's patterns match our prediction of delayed but statistically reliable conformity to the typical hearing children's six-step Guttman scale sequence, (b) replicability of the previously observed (Peterson et al., 2005) unique five-step sequence for autism and its capacity for extension to six Guttman scale steps, and (c) how individual task performance compares among these children and their peers with typical development. Finally, we tested in a more exploratory fashion whether children with AS progress in ToM understanding via the autism-specific sequence or the standard one, and how their overall rate of scale progress compares with that of children with autism on one hand, or typical development on the other.

### *Typically Developing Children*

The 68 typically developing children spanned a wide range of ToM understandings, from passing only one or two tasks (13%) to perfect performance on all six (26%). As shown in Table 1, fewer of them passed SARC (31%) than HE (53%), notwithstanding the procedural and conceptual similarities between these tasks that were outlined under Method. Comparing just these two tasks, there were 19 typically developing children who passed either SARC or HE, but not both. Seventeen of them (89%) passed HE only, whereas just 2 passed only SARC. The difference between these tasks was statistically significant, Wilcoxon signed-rank test: $z = 3.44$, $p = .001$, two-tailed.

Given this evidence for SARC's appropriate level of difficulty, we used Green's (1956) statistical procedures to assess conformity of typically developing children's response patterns to a perfect six-step Guttman scale in which any child who passes a harder task in the series will have passed each of the easier ones, and no child who fails any given task will have passed a later one. The responses of 59 of the 68 typical developers (87%) were perfectly consistent with the predicted six-step scale: DD > DB > KA > FB > HE > SARC. Green's index of reproducibility (Rep), assessing scale conformity (with values above .90 deemed significant) was .98. Green's index of consistency ($I$), a more stringent criterion that compares observed sequences with all that might be expected by chance (with values above .50 considered statistically significant) was .76.

Guttman analyses can be subject to skew when there are numerous items that most children in a sample pass. Even though our older children often did pass all the easier items, a further scalogram analysis for the typical group using only their responses to the four most difficult scale items (i.e., KA, FB, HE, and SARC; see Table 1) yielded strong evidence of scale reproducibility and consistency; 65 of 68 (96%) responded to the four advanced items in a manner that was perfectly scale consistent. Green's Rep was .99 and $I$ was .89, both significant.

Although SARC was clearly harder, it was not impossible: Twelve of the 29 (41%) oldest typical developers passed (see Table 1), as did 21 (31%) of the 68 typically developing children as a whole. Thus, the new task adds a further and more challenging scale step that allows charting progressive ToM growth across a broad age range (3–11 years). FB (ToM's traditional litmus criterion) occupies an intermediate point in this extended scale.

Age correlated significantly with total scale steps passed (ToM-6), $r(66) = .64$, $p < .001$, as well as with language scores on the CELF, $r(54) = .72$,

Table 1
*Means and Correct Responses on Key Measures by Diagnostic and Age Groupings*

| | Typically developing | | | | | |
|---|---|---|---|---|---|---|
| Diagnosis (group) | Oldest (Group 4A) | Middle (Group 4B) | Youngest (Group 4C) | Deafness (Group 1) | Autism (Group 2) | Asperger syndrome (Group 3) |
| *N* | 29 | 16 | 23 | 31 | 44 | 41 |
| Age (*SD*) | 8.77 (1.15) | 6.96 (0.40) | 4.75 (0.90) | 9.62 (1.70) | 9.02 (2.12) | 9.52 (2.16) |
| Mean CELF[a] (*SD*) | 21.85 (0.37) | 20.91 (1.22) | 19.53 (1.50) | 17.87 (2.99) | 17.45 (5.14) | 20.20 (2.55) |
| Pass DD | 29 (100%) | 14 (88%) | 22 (96%) | 29 (94%) | 41 (93%) | 40 (98%) |
| Pass DB | 29 (100%) | 15 (94%) | 19 (83%) | 29 (94%) | 38 (86%) | 38 (93%) |
| Pass KA | 29 (100%) | 16 (100%) | 17 (74%) | 20 (64%) | 31 (70%) | 36 (88%) |
| Pass FB | 29 (100%) | 11 (69%) | 10 (43%) | 16 (52%) | 19 (43%) | 26 (63%) |
| Pass HE | 23 (79%) | 8 (50%) | 5 (22%) | 6 (19%) | 23 (52%) | 26 (63%) |
| Pass SARC | 12 (41%) | 8 (50%) | 1 (4%) | 1 (3%) | 10 (23%) | 11 (27%) |
| Mean total (ToM-6)[b] (*SD*) | 5.21 (0.73) | 4.50 (1.46) | 3.22 (1.28) | 3.26 (1.36) | 3.68 (1.70) | 4.32 (1.47) |
| Mean total (ToM-5; *SD*) | 4.79 (0.42) | 4.00 (1.03) | 3.17 (1.19) | 3.23 (1.31) | 3.45 (1.40) | 4.05 (1.20) |

DD = diverse desires; DB = diverse beliefs; KA = knowledge access; FB = false belief; HE = hidden emotion; SARC = sarcasm understanding.

[a] Verbal ability (raw syntax score on Clinical Evaluation of Language Fundamentals–Preschool Test). [b] Total score on new six-step ToM Scale.

$p < .001$. CELF and ToM-6 were also significantly correlated, $r(54) = .65$, $p < .001$. With language ability partialled out, ToM-6 remained significantly correlated with age, $r(53) = .37$, $p < .01$, indicating that the new ToM Scale captured developmental progress on ToM understandings over and above any influence of language. Conversely, with age partialled out, CELF and ToM-6 scores also remained correlated, $r(53) = .33$, $p < .02$, highlighting the independent roles of both conceptual development—with age as its proxy here—and language development to typically developing children's progressive mastery of sequential steps in ToM understanding.

### *Comparisons of Typical Development With That of Children With Deafness, Autism, or AS*

This favorable developmental and scaling evidence underwrites comparisons of the overall pace of ToM Scale progress between typically and atypically developing groups. For these latter analyses, we selected only the oldest typically developing subgroup (Group 4A) to compare with children with deafness, autism, and AS since these four groups were well matched in age (see the Method section). Table 1 shows mean numbers of ToM-6 steps passed, as well as means on other variables including the CELF language ability test where a significant group difference, $F(3, 137) = 11.28$, $p < .001$, $\eta^2 = .20$, reflected Groups 3 and 4A equaling one another but significantly outperforming children with deafness and autism, who did not differ significantly.

Given this language ability contrast, we used analysis of covariance (ANCOVA) to compare ToM-6 scores across groups with chronological age and CELF raw scores as the covariates, $F(3, 135) = 5.75$, $p < .01$, $\eta^2 = .11$. Simple-effects planned comparisons showed that the typical developers in Group 4A scored significantly higher than children with AS, autism, and deafness (all *p*s < .02). When only the two focal delayed groups (children with deafness or with autism) were compared with one another in the same ANCOVA design, the difference did not achieve statistical significance, $F(1, 70) = 3.95$, $p > .05$, $\eta^2 = .05$, and a simple-effects planned comparison was likewise nonsignificant. Similarly, on the original five-step preschool scale, an ANCOVA with age and language scores covaried showed a significant overall difference among the four groups, $F(3, 135) = 4.96$, $p < .01$, $\eta^2 = .10$, which again reflected significantly higher scores for Group 4A than for each of the others (all *p*s < .02). Again, when children with autism versus deafness were contrasted via the same ANCOVA design, there was no significant difference, $F(1, 70) = 2.08$, $p > .15$, $\eta^2 = .03$, confirming a previous finding for different children from these same diagnostic groups (Peterson et al., 2005).

Thus, children with autism and with deafness were delayed relative to typically developing children on both the original (ToM-5) and extended scales (ToM-6). Importantly, these delays were apparent even after controlling statistically for age and language ability.

### *Independent Influences of Age, Language Skill, and Disability Type on Overall ToM Growth*

Just as for the typical developers (see above), simple correlations between chronological age and ToM-6 total scores were statistically significant for children with deafness, autism, and AS, $r(29) = .63$, $p < .001$; $r(42) = .48$, $p < .01$; and $r(39) = .47$, $p < .01$, respectively. The same was true of correlations between ToM-6 and language ability, $r(28) = .85$, $p < .001$; $r(42) = .33$, $p < .05$; and $r(39) = .75$, $p < .001$. We used a hierarchical regression analysis to more comprehensively examine the separate contributions of age and language ability to children's ToM-6 performance. Diagnostic group status was also included in order to see whether any of the specific developmental conditions we examined (deafness, autism or AS) made additional contributions over and above age and language skill to the total number of ToM steps children passed. The entire typically developing sample was included to optimize power and sensitivity for this analysis.

When CELF language scores were entered as the control variable at Step 1, they predicted ToM-6 scores as expected, $R = .55$, $R^2 = .30$, Adj. $R^2 = .30$, $F(1, 169) = 72.90$, $p < .001$. Importantly, the entry of chronological age as a separate predictor at Step 2 resulted in a statistically significant increment in the prediction, $R^2$ change = .06, $F$ change(1, 168) = 15.68, $p < .001$, and the full regression equation also remained significant, $R = .60$, $R^2 = .36$, Adj. $R^2 = .35$, $F(2, 168) = 47.46$, $p < .001$. Thus, even after controlling for language, children's progressive conceptual development (with age as its marker here) significantly contributed to overall progress through the ToM-6 Scale. At the final step, with the *en bloc* entry of diagnoses of deafness, autism, or AS (each dummy coded as 1 = present or 0 = absent, as recommended by Tabachnik & Fidell, 1996), there was a further significant increment in ToM-6 variability predicted, $R^2$ change = .07,

$F$ change (3, 165) = 6.43, $p < .001$. The full model was likewise significant, $F(5, 165) = 24.69$, $p < .001$; $R = .65$, $R^2 = .43$, Adj. $R^2 = .41$. Inspection of final beta weights indicated that the six separate predictor variables each made statistically significant contributions, all arising independently of one another: (a) age: $\beta = .44$, $t = 5.74$, $p < .001$; (b) deafness: $\beta = -.36$, $t = 4.38$, $p < .001$; (c) linguistic ability: $\beta = .32$, $t = 4.30$, $p < .001$; (d) autism: $\beta = -.24$, $t = 2.80$, $p < .01$; and (e) AS: $\beta = -.20$, $t = 2.64$, $p < .05$. In other words, not only did chronological age exert an independently predictive influence over and above linguistic maturity in the total number of ToM-6 steps that children passed, so did the independent influences of each type of developmental condition.

### *Scale Progressions for Focal Groups With Deafness or Autism*

These data on overall scale totals set the stage for examining key questions about developmental sequences of ToM understanding for our focal groups. Building upon Peterson et al.'s (2005) discovery of two distinct and reliable five-step ToM Scales, one characterizing deaf children (and typically developing preschoolers) and the other specific to children with autism, for the deaf children, we predicted statistically reliable conformity to the same six-step scale that was validated above for the present typically developing group (DD > DB > KA > FB > HE > SARC), whereas for those with autism, we predicted an alternative sequence, namely: DD > DB > KA > HE > FB > SARC where HE and FB steps were reversed.

Scale-consistent six-step sequences were observed for 29 of the 31 deaf children (94%). Green's Rep was .98 and $I$ was .75 (both significant) confirming the same six-step Guttman progression as for the typically developing children. It is worth comparing these scaling statistics to corresponding values obtained by Peterson et al. (2005) for a different sample of late-signing deaf children on the original 5-point ToM Scale. For those children, Rep was .95 and $I$ was .58. Thus, despite including an extra step, the present values compare very favorably. Note that 3, and only 3, deaf children in the present group (< 10%) had been previously included among the 36 late signers reported by Peterson et al. (2005), but they were tested at such different ages (11;3, 12;0, and 11;4 here vs. 5;5, 6;1, and 6;10 there) that their ToM understanding is unlikely to have remained the same. Just like typical developers, deaf children did worse on SARC than on HE, $z = 2.24$, $p < .05$, two-tailed Wilcoxon test.

Of the 44 children with autism, 35 (80%) displayed individual response patterns that were perfectly scale consistent with their alternative sequence (DD, DB, KA, HE, FB, SARC). Rep was .96 and $I$ was .67, both significant. For comparison, Rep for the autism-specific five-step scale in Peterson et al. (2005) was .95 and $I$ was .55 for their entirely separate sample. Thus, for children with autism or deafness, just as for the typically developing group, extending the scale to a further step of ToM-based sarcasm understanding provided a highly reliable, extended Guttman scale sequence. Pairwise task comparisons showed that SARC was passed significantly less often than HE by children with autism, $z = 3.61$, $p < .001$, two-tailed Wilcoxon test.

These data confirm a pattern from a previous study (Peterson et al., 2005) in which children with autism often passed the HE task out of the standard sequence (where FB precedes HE), while deaf and typical groups did not. To see if this specific contrast was replicable, we examined the numbers of children in each group who passed HE while failing FB, versus those displaying one of the three other patterns (i.e., passing FB not HE or failing or passing both tasks). No deaf child and only one typical developer (1%) did so, in contrast to nine children with autism (20%), $\chi^2(N = 143, df = 2) = 17.78$, $p < .001$.

Comparing patterns of individual task success and failure, the deaf and autistic groups both performed equally on four of the six tasks: $\chi^2(N = 75, df = 1)$ for DD, DB, KA, FB all < 1, all $p$s > .50. Those with autism did better on HE than both the deaf, $\chi^2(N = 75, df = 1) = 6.98$, $p < .01$, and typically developing preschoolers (Group 4C), $\chi^2(N = 67, df = 1) = 4.60$, $p < .05$ (all $p$s two-tailed, continuity corrected). Deaf children performed worse than even those with autism on SARC, $\chi^2(N = 75, df = 1) = 4.08$, $p < .05$.

### *Children With AS*

The children with AS performed no better than those with autism on any individual task, all $\chi^2(N = 85, df = 1) < 2.87$, all $p$s > .08. Furthermore, they were just as likely as those with autism to pass HE while failing FB, $\chi^2(N = 85, df = 1) = 1.14$, $p > .25$, and in this respect the two autism spectrum groups differed significantly from their peers in Groups 1 and 4, $\chi^2(N = 184, df = 3) = 16.80$, $p < .001$. Similarly, those with AS were no more likely than those with autism to pass FB while failing HE, $\chi^2(N = 85,\ df = 1) = .06$, $p > .75$. The AS group's mean ToM-6 score (see Table 1) was not significantly higher than the autism group's either before, $t(83) = 1.84$, $p = .07$, or after, $F(1, 81) < 1$, $p > .40$, $\eta^2 = .01$, controlling for age and CELF score differences via ANCOVA. Like each of the other groups, the children with AS found SARC significantly more difficult than HE, $z = 3.87$, $p < .001$, two-tailed Wilcoxon test.

With regard to scale progressions, in the absence of any previous evidence on ToM Scale sequencing for children with AS, we made no specific predictions but instead tested these children's conformity to each of the six-step sequences. Just as for those with autism, AS group's response patterns conformed to the distinctive six-step progression (HE > FB); 36 of the 41 children (88%) displayed response patterns that were perfectly consistent with this sequence and Rep = .97 and $I = .67$ were both statistically significant. However, although a numerically smaller proportion of children with AS (80% vs. 88%, *ns*) performed consistently with the typical developers' sequence (FB > HE) than with the autism one, Guttman scaling statistics (Rep = .97, $I = .66$) were significant for this order as well. Note that 7 children with AS had six-step scale patterns that were perfectly consistent with only one of these sequences, not both. Of these, 5 (71%) were consistent with the autism sequence only, compared to only 2 (28%) with the original sequence, $p > .20$, two-tailed binomial test.

## Discussion

These results provide several essential contributions. First and foremost, we validated a needed extension of the preschool ToM Scale (Wellman & Liu, 2004) so that it is now possible to chart the progress of ToM across six reliably sequential steps capturing a succession of changes, including those normatively arising in middle childhood, after FB is mastered. Our validation of this extended scale is particularly robust and useful because we provide data not only for younger and older typically developing children but also those with autism and deaf children of hearing parents. Second, the new scale's capacity to examine extended ToM trajectories for these two particular atypically developing groups is of special theoretical interest. As a single prefatory example, we confirm that the developmental sequence for children with autism is not only delayed but also uniquely different from that for other children. Third, by controlling statistically for individual differences in language ability, our study establishes that mastery of extended ToM progressions by typically and atypically developing children represents conceptual developments arising independently of increasing language competence alone—age predicted overall scale progress independently of linguistic maturity for all groups, with each specific disability type making additional independent contributions. No prior scaling study provides such data. Finally, the new scale enabled us to address, for the first time, a progression of ToM steps in children with AS, while comparing their progress to both autistic and typically developing groups.

### *An Extended ToM Scale That Adds an Understanding of Sarcasm*

A primary contribution of our study is the confirmation of a sixth, advanced step in the developmental sequence of ToM understanding. To create this extended scale, we added a task assessing understanding of nonliteral language, modified so as to conform to the structural and linguistic formats of the well-established preschool ToM Scale (Wellman & Liu, 2004). We focused on understanding irony or sarcasm owing partly to prior research (e.g., Filippova & Astington, 2008; Happé, 1994) suggesting that sophisticated ''mindreading'' (Baron-Cohen, 1995) is needed to genuinely understand it. Moreover, unlike many other nonliteral language forms, sarcastic, ironic interchange is a natural element in children's everyday conversation and social life. The results of our Guttman scaling analyses (which utilize a stringent approach to scaling; Festinger, 1947) confirmed the new sarcasm task's sequentiality with the other scale steps. At the same time, and as hoped, that task's requirement to understand discrepancies between spoken communicative intent and literal word meaning clearly made it substantially difficult for all groups of children. Specifically, it was more difficult than HE (previously the final scale step) and thus more difficult than understanding discrepancies between intended emotional expressions and true feelings. This sequential difference emerges despite close methodological and linguistic parallels between these tasks and despite their both reflecting familiar everyday aspects of children's social life requiring comprehension of the distinction between what is expressed versus what is felt. Yet, analyses also showed that while reliably more advanced in its cognitive ToM demands, the new sarcasm test was appropriately within the grasp of many older typically developing children, as well as at least some children with each of the other developmental conditions and that, even when they failed, the vast majority of children in all groups responded correctly to that task's control question.

This extended six-step scale seems a valuable addition to the ToM literature both psychometrically and theoretically. Psychometrically, it is valuable to have a reliable sequential index of developmental progress that extends measurement of ToM beyond (but also includes) normative, preschool achievements such as FB. This increases the potential age range for longitudinal (e.g., Wellman et al., 2011) and cross-cultural (e.g., Wellman et al., 2006) studies of social cognition in children with and without developmental delay. Moreover, the new scale offers fresh opportunities for fine-grained cross-sectional study of individual differences in ToM mastery, both within and between diagnostic groups. Understanding of nonliteral language clearly requires insight into the speaker's mind (Filippova & Astington, 2008) and sarcasm is superior to other more academic kinds of nonliteral language (e.g., simile, analogy, or metaphor) through being a natural element in children's everyday conversation and social life, and a well known problematic aspect of peer interaction for children autism or AS (e.g., Attwood, 2007) and for deaf children of hearing parents (e.g., Gregory et al., 1995).

In contrast, another commonly used test of advanced mental-state reasoning, the child version of the ''Reading the Mind in the Eyes'' test (Baron-Cohen, Wheelwright, Scahill, Spong, & Larson, 2001) is precluded from a developmental scale like the present one for several reasons. It includes 28 separate items, each in four-choice response formats that demand reading ages of about 10 years. Purely oral presentation to preliterate children (as were many in the present sample) is not feasible, requiring them to comprehend and remember four different emotion labels (e.g., ''ashamed,'' ''guilty,'' ''not believing,'' ''disgruntled'') per item (totaling 112 for the test as a whole) in order to make informed response choices. Indeed, Peterson and Slaughter (2009) found that these oral memory demands posed severe challenges even to adult university students, who displayed little response consistency when given the test in this purely oral manner.

Perner and Wimmer's (1985) second-order FB test, requiring recursive understanding of one story character's belief about another's belief, is another frequently used individual measure of older children's ToM. Like eye-reading, this task was also unsuitable for our scale partly owing to its methodology (e.g., a long and complex narrative with numerous essential test and control questions). More important, it would have been hard to interpret second-order task failure compared with failing our other scale items owing to the added executive, linguistic, and memory demands imposed by its procedure. Just as the eye-reading test requires literacy, the second-order task requires advanced domain-general executive functions (e.g., memory and planning), as well as advanced syntax simply to follow the procedure. These demands could extend the task's developmental trajectories, quite irrespective of its ToM demands. In fact, this non-ToM interpretation is suggested by Sullivan, Zaitchik, and Tager-Flusberg's (1994) finding that even 4-year-olds could pass a modified second-order task with multiple reminders, prompts, corrections, and story repeats to reduce its domain-general cognitive burdens. Dissanayake and Macintosh (2003) also found, paradoxically, that high-functioning children with autism did unexpectedly well on the second-order task despite scoring poorly on first-order FB (which should, in theory, have been easier). They suggested the heavy linguistic and memory burdens may have prompted chance success through guessing, an interpretation supported by high control-question failure rates.

Beyond these assessment issues and constraints, our expanded scale has theoretical grounding and implications. In devising their scale, Wellman and Liu (2004) conducted a meta-analysis of over 40 separate studies of typical preschoolers who took FB tests as well as a test of one other cognate ToM concept (e.g., obsolete desire, knowledge/ignorance). Those results informed the construction of the original scale. Conceptually, all of the subjective mental states that the scale focuses on require coordination between mentality and reality. This is equally true of the coordinated understanding of the objective situation (e.g., a rain-spoiled picnic) and the actual (literal) meaning of the speaker's words that our sarcasm task assesses. Thus, while well within the same mental-state domain, we showed that sarcasm was more difficult than either FB or emotional concealment for all groups. This can plausibly be interpreted, as noted earlier, in terms of the additional more complex coordinations of mind versus reality (Filippova & Astington, 2008) that an understanding of sarcasm requires.

We used single scale items to assess each sequential ToM concept, just as was done in the original scale, so that even this extended version could be completed by children, including those with developmental delay, in a single short session. Further research with multiple tasks for each scale step might be useful. However, complementary research provides assurance that our results are not limited to specific item characteristics. Meta-analyses of typically developing children's performance on our own and other formats of the DD, DB, and KA tasks (Wellman and Liu, 2004) and of FB (Wellman et al., 2001) confirm clear within-age consistencies, and clear cross-age differences, largely irrespective of story content, or task format. Wellman and colleagues (Wellman et al., 2006; Wellman et al., 2011) found that an alternative version of the HE task (about a boy wanting to conceal a negative emotional reaction to an adult's special gift) scaled in exactly the same way as the original story about peer teasing that we used. Similarly, Filippova and Astington (2008) used eight different sarcasm stories similar to ours and found all produced similar response patterns for typically developing children in the age range we tested, as did Happé (1995) with a pair of stories including her version of our SARC item. Furthermore, as noted in the Method section, our HE and SARC scenarios were very similar in both involving peers' affectively negative communication. Additionally, their use of parallel question formats extends the original scale's careful choice of comparable formats for all tasks.

### *Children With Autism or Deafness*

From a theoretical perspective, these new insights into ToM Scale progressions for children with autism and for deaf children of hearing parents are of special theoretical interest. In addition to these groups' widely confirmed FB delays (see Harris, 2006; Peterson, 2009; Tager-Flusberg, 2003, for reviews), their delayed development contrasts not only with the typical child's earlier timetable but also with the equally rapid scale progress (e.g., Peterson et al., 2005) and FB growth (e.g., Peterson & Siegal, 1999; Woolfe, Want, & Siegal, 2002) of natively signing children of deaf parents. Our new six-step assessment of ToM development confirmed further, extended delays for deaf children of hearing parents, and confirmed similar, extended delays for children with autism plus an alternative sequence, with comprehension of feigned emotion *preceding* FB.

*Scaling ToM development for children with autism*. How best to interpret this alternative sequence is not yet fully clear. One possibility is that the unique genetic and neurobiological features associated with the autism disorder (e.g., Tager-Flusberg, 2003) might be contributing factors. For example, understanding FB may constitute a representational ToM achievement of special neurocognitive processing difficulty, as is often assumed in neuroscience research on autism. However, it is also worth considering additional hypotheses, including that the data might reflect divergences in this groups' admittedly atypical social or conversational experiences. Conceivably, distinctive early socialization experiences, or educational training experiences, could have contributed to the confirmed reversal of HE and FB steps and to their overall delays. When reared in hearing families, deaf toddlers and preschoolers have little conversational exposure to others' mental states, especially to thoughts and beliefs not readily expressed via pointing or facial expression and, as we explain in detail below, this clearly contributes to their ToM delays behind their natively signing deaf peers. Following the same logic, the language delays that are integral to an autism diagnosis (as distinct from AS) may have the same restrictive effect on participation in mentalistic conversation. Indeed, naturalistic observation of parent–child interaction reveals a substantial reduction in discussing cognition when the child has autism (Slaughter, Peterson, & Macintosh, 2007). Impaired imagination may further curtail conversational exposure to others' belief states by precluding pretend play with siblings and peers. And children with autism typically participate in alternative educational situations and curricula, including explicit social-cognitive instruction. In sum, these distinctive experiences and restrictions could contribute both to ToM delays, and to unique task ordering, for children with autism.

Not only do these children's triadic linguistic-, social- and cognitive/imaginative-impairments (APA, 2000) lead to atypical early social and conversational experiences, but schooling places children with autism into a peer environment prone to frequent teasing (Attwood, 2007), including sarcasm or other ToM-dependent concepts. Speculatively, then, heightened social exposure to these peer group situations could boost cognitive reflection upon them by older children with autism. Relatedly, our HE task, by incorporating a scenario about peers' teasing, might sensitively probe the autism group's appropriate understanding of this particular ToM concept. To reiterate, however, this conversational-experiential hypothesis (see also Astington, 2004; Harris, 2006) is only one of several conceivable explanations for the autism-specific delays and reversal of scale steps and further research is now clearly warranted, ideally including direct observation of these children's peer interactions and any of the social skills education their teachers might offer.

*Deaf children's ToM Scale development*. The deaf children of hearing parents in our sample scored significantly below typically developing children on ToM-6 even after statistically controlling age and language ability. Indeed, as fully clear in Table 1, these older deaf children performed almost identically to the youngest typically developing group (3-, 4-, and 5-year-old preschoolers). In this way they paralleled delays seen in autism. Yet these late-signing deaf children were selected as being free not only of autism, or AS diagnoses, but also of any additional developmental conditions apart from hearing impairment. Furthermore, in the signing classrooms we recruited from, these children were not socially aloof and had extensive opportunities to interact socially and conversationally with signing deaf peers both in the classroom and on the playground.

No deaf native signers were included in the present sample (this 10% minority of the deaf population consists of the offspring of signing deaf parents acquire sign as their native language from birth rather than belatedly upon school entry). Past research (see Peterson, 2009, for a review) indicates that native signers are swifter than late signers to master both FB (e.g., Courtin & Melot, 1998; Peterson & Siegal, 1999; Schick, de Villiers, de Villiers, & Hoffmeister, 2007; Woolfe et al., 2002) and the full five-step ToM Scale (Peterson et al., 2005). Furthermore, this ToM contrast between native signers and late signers is found to apply over and above any contrasts in signed or spoken language skill (Woolfe et al., 2002). It is likely that late signers' restricted access to early family conversation at home accounts, at least in part, for their slow ToM trajectory.

While this will explain overall pace of delays on the six-step ToM Scale displayed by the deaf late signers in our sample, their unusually poor performance on the sarcasm task still remains remarkable. Even though sarcasm was, as expected, the hardest of all the tasks for every group, it was harder still for deaf children even than for their peers with autism. Indeed, despite their equal overall ToM-5 and ToM-6 scores, there was a sevenfold difference in proportions of deaf versus autism groups who passed sarcasm (3% vs. 23%). What could explain the deaf group's unusual difficulties? We know of no previous empirical studies of deaf children's comprehension of irony or sarcasm, but one early test of deaf adolescents' identifications of the figurative meanings of idioms (e.g., by pointing at a happy face for ''She is over the moon'') showed that only those with an advanced reading age scored above chance (Fruchter, Wilbur, & Fraser, 1984). Similarly, case histories (e.g., Gregory et al., 1995) indicate that conversational sarcasm and word play often continue to confuse deaf adults, irrespective of their mature fluency in sign or speech. Thus, perhaps early family conversation and/or extended language experiences are important here as well. In the signing classrooms we recruited from, deaf children (with a peer group of other deaf signers) are unlikely to encounter sarcasm, since this particular form of humor is rarely used even by deaf teenagers and young adults (Gregory et al., 1995). Obviously, further research into deaf children's understanding of sarcasm is now warranted, together with a further examination of the promising links shown here between this understanding and that of other sequential ToM components.

*ToM development in AS*. Though less central to our primary theoretical and methodological aims, our novel inclusion of a group of children with AS was helpful for clarifying and extending the mixed results of previous studies of: (a) whether ToM is delayed for children with AS relative to the typical timetable and (b) comparative rates of ToM mastery for children with AS versus high-functioning autism. In much of this past research only a single type of task has been used (usually FB), making it hard to distinguish task-specific from general ToM-related contrasts. Focally, results of past studies are inconsistent with one another. As noted in the Introduction, for example, Paynter and Peterson (2010) found that even after controlling for age, IQ, and language ability (vocabulary plus syntax), children with AS scored as highly as matched typical developers on a FB battery and above their peers with autism. Yet Dissanayake and Macintosh (2003) used a similar FB battery and found that younger typically developing children outperformed older children with AS. With comprehensive evidence across six sequentially connected tasks, we show that children with AS were indeed slower to master overall ToM than their peers with typical development, even after controlling for age and language ability. Furthermore, children with AS and autism performed equivalently once the effects of age and language were statistically controlled.

These patterns have interesting implications. Recall that an AS diagnosis via the *DSM–IV* ''gold standard'' (applied stringently in our study) implies that although they diverge from their peers with autism on one core symptom (early language delays), children with AS and autism converge on two core symptoms (debilitating deficits in socialization and imaginative/cognitive flexibility; Szatmari et al., 2009). In our study, despite still displaying this language discrepancy, these two groups were equivalently delayed in overall ToM Scale progress. Thus, in terms of delayed ToM understanding, the distinguishing language feature appears less relevant than these groups' two shared diagnostic impairments. Perhaps, following the above line about likely influences of atypical early peer and family socialization for children with autism, similarly atypical early social experiences consequent upon these shared diagnostic symptoms might slow the growth of ToM for children with AS. At school, these children are frequently victimized by peer exclusion, taunting or ridicule (Attwood, 2007). Furthermore, their more advanced language skills could increase their sensitivity to, or confusion by, the ToM paradoxes entailed by sarcastic teasing or false emotional displays.

Thus, in line with reasoning articulated above for autism, heightened socialization pressures may conceivably have elevated the AS group's ToM performance specifically on SARC and HE tasks, just as we speculated for autism. This interpretation is broadly in keeping with Szatmari et al.'s (2000) finding that social immaturity persists in children with AS even in the context of clinically normal language skills. While not assessing ToM, these authors followed a group of children with AS longitudinally from age 4 to age 7. Despite having had no initial delays in language onset (consistent with their diagnosis) and despite maintaining normal levels of language maturity throughout early and later childhood, these children scored as significantly less socially mature than their typically developing peers throughout the study. Furthermore, in the same study, children with autism who had largely ''caught up''(p. 1986) linguistically by age 7 to their peers with AS nevertheless remained substantially delayed in social maturity independent of language skill and nonverbal IQ. In other words, serious peer relations problems are likely to persist well into middle childhood in groups with AS and autism even when these children's language skills are not seriously impaired. Such immaturely atypical peer social relations might therefore conceivably relate to the delays in ToM understanding that we observed in both of these groups with autism spectrum disorders in the manner suggested above.

At the same time, the scale performance of our samples with autism and AS may help to diminish earlier concerns that non-ToM-related heuristics or ''hacking'' strategies (Happé, 1995) could enable some language-adept children with these developmental conditions to pass FB tests even when they have no genuine understanding of ToM. Hacking alternatives to mental-state reasoning are believed to be cumbersome, time consuming, and task specific (see Happé, 1995). Furthermore, these heuristics apply to the characteristic features specific to one task or another; most currently formulated hacking heuristics are specific to FB. But our results show more comprehensive delays and peculiarities, *and* more comprehensive developments. Indeed, as clear in Table 1, the same children with autism or AS who answer FB correctly are as good, *or better* at understanding HE. Arguably, something more than low-level hacking seems to characterize the ToM gains these children developmentally achieve progressively across all six scale steps.

### *Conclusions*

The new six-step ToM Scale provides a valid, more extended sequential progression of ToM insights applicable with younger children but now also with older more advanced children beyond the preschool range. It should be useful for providing more encompassing and efficient evaluation of levels and progressions of ToM development in typically and atypically developing groups. The extended sequence also has clear value for the theoretical challenges of explaining how ToM develops in children generally. In particular, these findings include new evidence about sarcasm's place in the ToM progression, amplifying data and ideas about the role of conversational and socially interactive experience (Astington, 2001; Harris, 2006) in ToM growth. Such experiences likely contribute both to social-cognitive advances and to special challenges encountered by some intelligent but atypically developing children. Active and varied participation in social exchanges (Harris, 2006; Wellman, 2002) is likely to assist all children's timely achievement of mature social understanding. But, with development, these interchanges gain complexity and increasingly include opinion exchange and concealment, shared fantasizing, angry belief-based disputes, teasing, joking, sarcasm, and other affectively laden nonliteral uses of language. Even though such conversational exchanges are part and parcel of everyday social interaction, some of them clearly continue to pose challenges that extend well into middle childhood, especially for children with high-functioning autism or deafness in a hearing family.

ToM is a developmental phenomenon now shown to progress in a statistically reliable sequence of steps beyond preschool. This is true both for typically children and those with ToM delays owing to deafness or autism-spectrum disorders. This extended development has been surprisingly little studied. The six-step ToM Scale validated here can help us better understand the comprehensive nature of ToM development. It does so in the current research, and could do so in further applications to typically developing and atypically developing children.

## References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Astington, J. W. (2001). The future of theory-of-mind research: Understanding motivational states, the role of language and real-world consequences. *Child Development*, *72*, 685–687.

Astington, J. W. (2004). What's new about social construction? Distinct roles needed for language and communication. *Behavioral & Brain Sciences*, *27*, 96–97.

Attwood, T. (2007). *The complete guide to Asperger syndrome*. New York: Jessica Kingsley.

Banerjee, R. (2000). Developing an understanding of modesty. *British Journal of Developmental Psychology*, *18*, 499–517.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Baron-Cohen, S. (2000). Theory of mind and autism. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds* (pp. 3–17). Cambridge, UK: Cambridge University Press.

Baron-Cohen, S., Wheelwright, S., Scahill, V., Spong, A., & Larson, J. (2001). Are intuitive physics and intuitive psychology independent? *Journal of Development & Learning Disorders*, *5*, 47–78.

Courtin, C., & Melot, A. (1998). Development of theories of mind in deaf children. In M. Marschark & M. D. Clark (Eds.), *Psychological perspectives on deafness* (Vol. 2, pp. 79–102). Mahwah, NJ: Erlbaum.

Dahlgren, S., & Trillingsgaard, A. (1996). Theory of mind in nonretarded children with autism and Asperger syndrome: A research note. *Journal of Child Psychology & Psychiatry*, *37*, 759–763.

Dissanayake, C., & Macintosh, K. (2003). Children with autistic disorder and Asperger's disorder. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind* (pp. 213–240). New York: Psychology Press.

Festinger, L. (1947). The treatment of qualitative data by scale analysis. *Psychological Bulletin*, *44*, 149–161.

Filippova, E., & Astington, J. W. (2008). Further development in social reasoning revealed in discourse irony understanding. *Child Development*, *79*, 126–138.

Fruchter, A., Wilbur, R., & Fraser, J. (1984). Comprehending idioms by hearing-impaired adolescents. *Volta Review*, *86*, 7–19.

Green, B. F. (1956). A method of scalogram analysis using summary statistics. *Psychometrika*, *21*, 79–88.

Gregory, S., Bishop, J., & Sheldon, L. (1995). *Deaf young people and their families*. Cambridge, UK: Cambridge University Press.

Happé, F. (1994). An advanced test of theory of mind. *Journal of Autism & Developmental Disorders*, *24*, 129–154.

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

Harris, P. L. (2006). Social cognition. In W. Damon (Ed.), *Handbook of child psychology* (pp. 811–857). New York: Wiley.

Jolliffe, T., & Baron-Cohen, S. (1999). The strange stories test: A replication with high-functioning adults with autism or Asperger syndrome. *Journal of Autism and Developmental Disorders*, *29*, 395–406.

Kristen, S., Thoermer, C., Hofer, T., Aschersleben, G., & Sodian, B. (2006). Skalierung von Theory of Mind Aufgaben [Scaling of theory of mind tasks]. *Zeitschrift für Entwick -lungspsychologie und Pädagogische Psychologie*, *38*, 186–195.

Ozonoff, S., Rogers, S., & Pennington, B. (1991). Asperger's Syndrome: Evidence of an empirical distinction from high-functioning autism. *Journal of Child Psychiatry and Psychology*, *32*, 1107–1122.

Paynter, J., & Peterson, C. C. (2010). Language and theory of mind development in autism versus Asperger's syndrome. *Research in Autism Spectrum Disorders*, *4*, 377–385.

Perner, J., & Wimmer, H. (1985). ''John thinks that Mary thinks that . . .'' *Journal of Experimental Child Psychology*, *39*, 437–471.

Peterson, C. C. (2009). The development of social-cognitive and communication skills in children born deaf. *Scandinavian Journal of Psychology*, *50*, 475–483.

Peterson, C., & Siegal, M. (1999). Representing inner worlds: Theory of mind in autistic, deaf, and normal hearing children. *Psychological Science*, *10*, 126–129.

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

Peterson, C., Slaughter, V., & Paynter, J. (2007). Social maturity and theory of mind in typically developing children and those on the autism spectrum. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *48*, 1243–1250.

Peterson, C. C., & Wellman, H. M. (2009). From fancy to reason: Scaling deaf children's theory of mind and pretence. *British Journal of Developmental Psychology*, *27*, 297–310.

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

Pons, F., Lawson, J., Harris, P., & deRosnay, M. (2003). Individual differences in emotion understanding. *Scandinavian Journal of Psychology*, *44*, 347–353.

Rajendran, G., Mitchell, P., & Rickards, H. (2005). How do individuals with Asperger syndrome respond to nonliteral language and inappropriate requests in computer-mediated communication. *Journal of Autism and Developmental Disorders*, *35*, 429–443.

Remmel, E., & Peters, K. (2009). Theory of mind and language in children with cochlear implants. *Journal of Deaf Studies and Deaf Education*, *14*, 218–236.

Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and theory of mind understanding. *Child Development*, *73*, 734–751.

Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and theory of mind: A study of deaf children. *Child Development*, *78*, 376–396.

Sharp, C., & Fonagy, P. (2008). Social cognition and attachment-related disorders. In C. Sharp, P. Fonagy, & I. Goodyer (Eds.), *Social cognition and developmental psychopathology* (pp. 271–302). New York: Oxford University Press.

Siegal, M., & Peterson, C. C. (2008). Language and theory of mind in atypically developing children. In C. Sharp, P. Fonagy, & I. Goodyer (Eds.), *Social cognition and developmental psychopathology* (pp. 81–112). New York: Oxford University Press.

Slaughter, V., Peterson, C., & Macintosh, E. (2007). Mind what mother says: Narrative input and ToM in typical children and those on the autism spectrum. *Child Development*, *78*, 839–858.

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

Szatmari, P., Bryson, S., Duku, E., Vaccarella, L., Zwaigenbaum, L., Bennett, T., et al. (2009). Similar developmental trajectories in autism and Asperger syndrome from early childhood to adolescence. *Journal of Child Psychology and Psychiatry*, *50*, 1459–1467.

Szatmari, P., Bryson, S., Streiner, D., Wilson, F., Archer, L., & Ryerse, C. (2000). Two-year outcome of preschool children with autism or Asperger's syndrome. *American Journal of Psychiatry*, *157*, 1980–1987.

Tabachnick, B., & Fidell, L. (1996). *Using multivariate statistics*. New York: HarperCollins.

Tager-Flusberg, H. (2003). Evaluating the theory-of-mind hypothesis of autism. *Current Directions in Psychological Science*, *16*, 311–315.

Wellman, H. M. (2002). Understanding the psychological world: Developing a theory of mind. In U. Goswami (Ed.), *Handbook of childhood cognitive development* (pp. 167–187). Oxford: Blackwell.

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

Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science*, *17*, 1075–1081.

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory of mind scale: Longitudinal perspectives. *Child Development*, *82*, 780–792.

Wellman, H. M., & Liu, D. (2004). Scaling theory of mind tasks. *Child Development*, *75*, 759–763.

Wiig, E., Secord, W., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals (CELF)–Preschool*. San Antonio, TX: Psychology Corporation.

Woolfe, T., Want, S., & Siegal, M. (2002). Signposts to development. *Child Development*, *73*, 768–778.

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 typically developing individuals. *Psychological Bulletin*, *124*, 283–307.

## Appendix

### Examples of Scored Responses to the SARC Test Question (*''Why did she say 'It's a lovely day'?''*) Taken From Each Group's Transcripts

| Pass (*shows explicit awareness of discrepancy between literal and intended meaning*) | Fail (*no evidence of perceiving nonliteral meaning*) |
|---|---|
| She's being sarcastic/sarcasm | Because it is sunny [raining] |
| She doesn't mean it | It's lovely outside |
| Because it's an idiom | She likes rain [picnics]/We need rain |
| She tricked him | She wants to play in the puddles |
| Its her way of telling him she is upset | Because she [the cake] got wet |
| Just to make up a little joke | Because he lied to her |
| She is saying politely that she is not happy | She thought it was sunny/did not see clouds |
| Because she is a smart aleck | She's cross |
| Because she is meaning ''Why tell me it was nice?'' | To tell him off |
| | Because it's not sunny |
| | So he doesn't feel bad |
| | Because he said it first |
| | Because her Dad likes the rain but she doesn't |