Article type : Empirical Article

Longitudinal Theory of Mind (ToM) Development from Preschool to Adolescence with and without ToM Delay

Candida C. Peterson & Henry M. Wellman

University of Queensland University of Michigan

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/cdev.13064

*Corresponding Author:
Professor Candida C. Peterson, School of Psychology,
University of Queensland, Brisbane, Queensland, Australia 4072
email: candi@psy.uq.edu.au; fax: 617-3365-4466.

Running Head: Longitudinal ToM

Abstract

Longitudinal tracking of 107 3- to-13-year-olds in a cross-sequential design showed a 6-step theory of mind (ToM) sequence identified by a few past cross-sectional studies validly depicted longitudinal ToM development from early to middle childhood for typically developing (TD) children and those with ToM delays owing to deafness or autism. Substantively, all groups showed ToM progress throughout middle childhood. Atypical development was more extended and began and ended at lower levels than for TD children. Yet most children in all groups progressed over the study's mean 1.5 years. Findings help resolve theoretical debates about ToM development for children with and without delay and gain strength and weight via their applicability to three disparate groups varying in ToM timing and sequencing.

Longitudinal Theory of Mind Development from Preschool to Adolescence with and without Delay

Theory of mind (ToM), or the representational understanding of how thoughts and feelings shape human behavior, is fundamental to social life and social reasoning. Sometimes dubbed the "quintessential ability that makes us human" (Baron-Cohen, 2001, p. 174), ToM understanding is a correlate of children's everyday social concerns over friendship (Fink et al., 2015), popularity (Slaughter, Imuta, Peterson & Henry, 2015), leadership (Peterson, O'Reilly & Wellman, 2016) and

loneliness (Devine & Hughes, 2013), as well as with persuasion (Slaughter, Peterson & Moore, 2011) and deception (e.g., Ding et al. 2015) skills. Since poor social perception and problematic peer relations pose adverse risks for mental health throughout the remainder of life (Bagwell, Newcomb & Bukowski, 1998) the development of ToM understanding through early and middle childhood has both practical and theoretical significance.

Too little is known about this extended developmental trajectory, partly because empirical documentation of later ToM developments is scanty. For example, based on a Scopus search of more than 6,000 published ToM studies, Hughes (2016) found less than 4% included school-aged children, prompting calls for research to map ToM through the "uncharted waters of middle childhood" (p. 4). Promising research with typically developing (TD) youth for this post-preschool period is emerging (Devine & Hughes, 2012; Miller, 2009), but studies comparing typical developers and those with ToM delay seem particularly promising and needed. Unlike typical development, certain groups, such as deaf children from hearing families (DoH children) and children with autism, are often substantially and selectively delayed in ToM mastery (for reviews see Baron-Cohen, 1995, 2001; Happé, 1995; Peterson, 2009; Siegal & Peterson, 2008), routinely failing preschool false belief tests throughout middle childhood and even the teens. Documentation of these delays is abundant and clear, but important questions remain as to their nature and basis.

With this background, we next consider the theoretical motivations of our study, including (a) new insights to be gained from focusing on a developmental sequence of ToM transitions across the whole of childhood, (b) the added explanatory power of doing so longitudinally via a cross-sequential design and (c) the particular theoretical reasons to include groups of children with autism or deafness. In this theoretical context, methodologies for measuring and comparing typical and atypical children's ToM progress over this broad age range are examined.

Theoretical Questions about Longitudinal ToM Development

Typical ToM growth. Assuming an informative yardstick for measuring school-aged children's ToM progress, key theoretical questions arise. One is whether ToM development exhibits a variable versus steady change across development. For example, how does ToM growth compare between younger (aged 3 to 5) and older (aged 7 to 13) children? Possibly ToM gains may taper off at the end of preschool, once the false belief milestone is mastered. Alternatively, immersion in the school-child's new peer-oriented social world could trigger a spurt in ToM growth, escalating with the ever more intricate social challenges preceding adolescence.

Questions about extended comparative ToM change have been difficult for past studies to answer, even the few employing a longitudinal methodology. Some past studies have focused exclusively on a single ToM concept for all testings (e.g., first-order false belief: Razza & Blair, 2009; or faux pas: Banerjee, Watling & Caputi, 2012). Others have used completely different ToM

measures at each test time (e.g., false belief at Time 1 but Strange Stories at Time 2: Devine, White, Ensor & Hughes 2016). These approaches limit the kinds of conclusions that can be drawn across development. One theoretical question, for example, concerns the prevalence and stability throughout childhood of individual differences. These exist for TD preschoolers (Dunn, 1995) and their correlates have long been of theoretical interest. What happens later is less clear. Devine et al. (2016) wondered whether the individual difference phenomenon is unique to early childhood. Their longitudinal findings for TD 6- to 10-year-olds showed this was not so, thus "extending the current literature by documenting the developmental reach of individual differences in ToM" (p. 766). Devine et al. (2016) proposed two contrasting hypotheses. One, that we term an "even gains" model, predicts that individual differences among typical developers may maintain themselves at a constant rate throughout development as each new ToM concept is mastered. We follow Bornstein et al. (2014) who labeled this "consistency in individual differences over time" as "stability" (p. 1346) and found that it characterized children's core language skills from age 4 to 14 years. An alternative possibility for ToM is that a "catch-up" might occur such that TD children who are initially at the slower end of normal development might later accelerate to equal or overtake TD peers who achieved their first ToM steps exceptionally early.

ASD and DoH children. These theoretical questions and others arise even more so for children with ToM delay. Much research shows that children with autism spectrum disorder (ASD) differ informatively from TD groups in the timing of their ToM mastery (e.g., Baron-Cohen, Leslie & Frith, 1985; Happé, 1995). The same is true of many severely or profoundly deaf children of hearing parents (DoH children) when they grow up in hearing-only households where, initially at least, no one has sufficient signing fluency to communicate freely about cognitive and affective mental states (e.g., Courtin & Melot, 1998; Meristo et al. 2007; Peterson & Siegal 1999; Schick et al., 2007; Vaccari & Marschark, 1997). However, here too, a lack of crucial longitudinal data precludes addressing questions like (a) constant versus variable trajectories over the transition to school, (b) within-group catch-up or stability and (c) individual variability in the possibility for, and extent of longitudinal progress (e.g., do most DoH or ASD children make ToM progress, or only a small minority?).

Of course, it is important to note that "catching-up" here is relative. Given the substantial evidence of severe ToM delay persisting through adolescence into adulthood for ASD and DoH individuals (e.g., Holroyd & Baron-Cohen, 1993; Pyers & Senghas, 2009), we do not predict any catching up to equal TD groups. However, within any ASD or DoH group there are likely to be individual differences. Not all children with the same disability will master a particular ToM concept at identical age. Knowledge of these within-group individual differences could prove theoretically illuminating because, for example, if post-preschool longitudinal progress is evident

for many or most of the ASD and DoH children who frequently play and converse with peers during middle childhood, this could implicate these experiences in ToM development. Conversely, if little or no longitudinal gain is evident for any children in these groups despite their active conversational participation, such a theory could prove less tenable.

Sensitive periods? At the extreme**,** questions about the presence versus absence of longitudinal gain merge into discussions of whether certain periods in development are more conducive than others to ToM growth. As an analogy, for neurological development, Tierney and Nelson (2009) noted that "experiences in the early years of development can affect the development of the brain in ways that later experiences do not" (p. 11). They explained this in terms of an "experience-expectant sensitive period", defined as a time when the human brain is biologically primed to be optimally ready for certain developmentally significant experiences. Deprivation of these expected experiences during this sensitive time can have lasting adverse consequences. For ToM, the years before school entry have been suggested as such a period. For example, Siegal and Varley (2002) wrote: "These findings point to a critical period in ToM. Just as children seem to be irreparably impaired in their later language learning when not exposed at all to an early language environment, children require at least some minimal access to conversation about mental states to show ToM reasoning" (p. 469). We use the term "sensitive period" in place of Siegal and Varley's "critical period" to reflect the more moderate view that after the period's end there is "reduced but not absent capacity to learn" (Newport, 1991. p. 739).

In contrast to sensitive periods, an alternative theory proposes"experience dependent plasticity" (Tierney & Nelson, 2009). On this theory, there is no one period in the lifespan that is more sensitive than any other to relevant experience for the development of a focal capacity. For this alternative, ToM growth is feasible at any age, including adulthood, provided the necessary stimulating experiences (like mentalistic conversation) eventually occur.

Longitudinal data (as opposed to cross-sectional comparisons of different people of different ages) are needed to address questions about the existence or non-existence of sensitive periods for ToM. Yet longitudinal studies, especially of ASD and DoH groups in late childhood and the early teens are rare. Nonetheless, two past longitudinal studies of autistic individuals' false belief performance (Holroyd & Baron-Cohen, 1993; Ozonoff & McEvoy, 1994) have reached provocative conclusions. No longitudinal gains in false belief understanding emerged even in the teens or adulthood in either study, despite longitudinal intervals of 3 to 8 years. While these results might seem to support limited post-preschool growth conclusions, and even experience-expectant sensitive periods, the opposing alternative of experience-dependent plasticity cannot be ruled out because no information was given about participants' social situations. Furthermore, Howlin, Mawhood and Rutter's (2000) longitudinal follow-up (from age 7 to 24 years) of the social

behavior of ASD individuals from an era and background similar to Holroyd and Baron-Cohen's sample suggests that even in adulthood access to requisite ToM-stimulating experiences may have been very limited. Although ToM was not measured, most (e.g., 89%) were rated by clinicians as having no friends and too few social skills for normal social participation, thus limiting even belated access to the kinds of social and conversational inputs that can experientially nurture ToM (Harris, 2005).

On the other hand, longitudinal data for a small sample of DoH adults in Nicaragua (Pyers & Senghas, 2009) were more consistent with the experience-dependent plasticity alternative. Eight DoH adults who had had little or no access to sign language in childhood and had failed first-order false belief when first tested at mean age 26.8 years were retested approximately two years later. Significant longitudinal ToM gains were made, coinciding with participation in a deaf social club that gave them access to mentalistic conversation with other adult signers.

These provocative findings for ASD and DoH individuals are of course inconsistent across studies and sample populations. Moreover, sample sizes were uniformly small and ToM was measured by false belief measures alone (mostly in multi-item false belief batteries) rather than a developmentally-sequenced ToM Scale.

Developmental ToM sequences. Theoretically, ToM as a cornerstone of social cognition and social competence is not limited to false belief or to the preschool period. Thus longitudinal sequences of ToM growth are of special theoretical interest. ToM defined broadly includes children's intuitions about the nature and behavioral consequences of mental states in general, not just false belief (e.g., Astington, 2001). Even toddlers and young preschoolers who do not yet understand false belief are found to have such intuitions. Indeed, some have claimed (e.g., Onishi & Baillargeon, 2005), based on passive nonverbal looking-time procedures, that infants as young as 7 to 12 months possess "implicit" ToM concepts. However such claims are controversial. Others have argued that infant looking-time data can be explained more parsimoniously via simpler, non-mentalistic awareness of behavioral regularities or visual novelty (e.g., Heyes, 2014; Ruffman & Perner, 2005). Nevertheless, irrespective of these debates about ToM in infancy, there is ample evidence from "standard" preschool tests requiring explicit judgments and active behavioral choices that toddlers (aged 18-24 months) and young preschoolers (aged 2 to 3 years) understand at least some pre-false-belief ToM concepts like desires, intentions and/or true beliefs (e.g., Gopnik & Slaughter, 1991; Meltzoff, 1995; O'Reilly, J. & Peterson, 2014; Poulin-Dubois & Yott, 2017; Wellman & Woolley, 1990: see also the meta-analysis by Wellman & Liu, 2004).

Conversely, some aspects of ToM develop well after false belief and can challenge even adults (e.g., sarcasm: O'Reilly, K. et al., 2013). Thus it is clearly of theoretical interest to explore the nature and extent of longitudinal ToM growth up to and beyond false belief, and how these

concepts are sequenced. For example, delayed ToM growth in children with autism or deafness could either map onto a "standard" sequential trajectory from preliminary (e.g., desire) to more advanced (e.g., false belief) ToM concepts. Alternatively, the longitudinal sequences of ToM milestones for TD, ASD and DoH children could be different. Potentially, for ASD and DoH groups, restricted conversational access to others' mental states (e.g., see the social-communicative cascade account of ASD put forth by Mundy, Sullivan & Mastergeorge, 2009) could undermine ToM development. Alternatively, the root causes for delay could be different for each group (for example an innately intractable autism-specific neuro-cognitive deficit in ASD (Leslie & Thaiss, 1994) versus restricted conversation for the deaf). If so, then the capacity for post-preschool longitudinal growth could differ for ASD and DoH groups, despite equivalent delay. Further investigation of longitudinal growth in both typical and atypical ASD and DoH children is clearly needed, and in particular comparison of ASD, DoH and TD groups on the same measures within the same study.

Measuring ToM Progress

One issue constraining progress in answering theoretical questions like these is how best to empirically measure children's ToM understanding. For preschoolers, the answer may seem clear. One often-used, well-validated "litmus" ToM test, assessing explicit false belief understanding, has been the overwhelming choice for studies of thousands of children over several decades. Most 3-year-olds consistently fail. Yet, by age 5 to 6, consistent success is achieved near-universally by typically developing (TD) children in many cultures worldwide (e.g., Callaghan et al., 2005; Wellman, Cross & Watson, 2001). Note that only about 3 years of development separate TD children's consistently floor versus ceiling performance, resulting in a narrow age window for studying development and individual variability. How can ToM development be measured in older children who already score at ceiling on standard false belief tests?

An initial proposal, continuing a focus on false belief, was to attempt to develop a "second-order false belief" procedure to examine understanding that someone may hold a false belief about another person's belief (Perner & Wimmer, 1985). Despite passing first-order tasks, TD children often fail second-order false belief as late as age 7 to 9. However, it is unclear whether such failure reflects limited understanding in the ToM domain versus lack of more general skills for memory, syntax or executive functioning. Some cross-sectional evidence suggests the latter (e.g., Sullivan, Zaitchik & Tager-Flusberg, 1994; Tager-Flusberg & Sullivan, 1994).

An alternative, pioneered by Wellman and Liu (2004), uses statistical scaling methods (Guttman and Rasch analyses) to construct multi-item measures, or scales, representing extended ToM growth as a series of increasingly difficult levels all sharing the same unified conceptual basis (for other examples see Osterhaus, Koerber & Sodian, 2016; Pons, Harris & de Rosnay, 2004).

Wellman and Liu's (2004) 5-step ToM Scale is designed for preschoolers aged 2 to 6 years and comprises five ToM tasks, assessing a continuum of preschool ToM concepts beginning with diversity of people's desires and proceeding through true belief to false belief and hidden emotion. Widespread cross-sectional data reveal that patterns of individual task success and failure define a Guttman sequence where it is rare to pass the next scale step without also passing the preceding one. One longitudinal study (Wellman, Fang & Peterson, 2011) demonstrated that the 5-step sequence also characterized individual children's progress over time. The problem with applying this scale to the "uncharted" (Hughes, 2016) period of middle childhood is ceiling performance by most TD children over age 5. For example, in Wellman et al.'s (2011) study, most TD children in the US and China passed all five scale tasks at mean ages of 5.5 and 5.3 years, respectively.

The 6-Step ToM Scale. Peterson, Wellman and Slaughter (2012) added a further more advanced step to the Wellman and Liu (2004) scale, providing an extended 6-step ToM Scale comprising six discrete ToM concepts. Cross-sectional evidence (e.g., Peterson et al. 2012; Peterson, O'Reilly & Wellman, 2016; Peterson, Slaughter, Moore & Wellman, 2015) has confirmed not only the 6-step scale's reliable Guttman sequentiality, but also that the final two scale steps continue to prove highly challenging for older TD children in the age range 7 to 12 years. Thus, in Peterson et al. (2016) 67% of TD children failed one or both of these at ages 8 to 11 years. Consequently, the 6-step scale holds promise but requires longitudinal validation, given that all the (limited) past evidence for it has been cross-sectional.

Longitudinal designs for ToM. Patterns of success and failure by different age groups in a cross-sectional study do not unequivocally identify developmental change, providing only indirect proxies for age-related change and, indeed, can arise for reasons having nothing to do with development per se (Schaie, 1972). Longitudinal studies are needed. Cross-sequential longitudinal designs (Schaie, 1972), by longitudinally tracking a complete cross-sectional sample over time, can be particularly revealing, extending beyond limits of cross-sectional or longitudinal data when used alone. For example, Schaie (1972) demonstrated that cross-sequential IQ data from 3 cohorts of first-grade boys (longitudinally tracked for 4 months) yielded novel information unavailable from either the simple cross-sectional or simple longitudinal components treated separately.

The Present Research

We supply cross-sequential longitudinal data via the 6-step ToM Scale to address substantive questions of significance for theories of ToM development. By including ASD and DoH children, our data can confirm the longitudinal validity of the 6-step scale in three carefully contrasting groups and can begin to explore theoretical questions about the nature, antecedents and consequences of ToM progress in typical and atypical development, as well as about ToM sequences. For example, our longitudinal use of an extended developmentally-sequenced ToM

Scale enables comparative exploration for ASD, DoH and TD children of ToM sequences and timetables over a broad age range, thereby clarifying ambiguities and inconsistencies in past cross-sectional research. To illustrate, Wellman et al.'s (2011) longitudinal use of the preschool (5-step) revealed DoH children's delays at each scale step, not just false belief. But that study only examined preschool ToM concepts. Plus no corresponding data exists at all for ASD groups given no previous longitudinal evaluation of their ToM Scale performance.

Method

Participants

Three groups totaling 107 Australian children (Time 1 age: 3 to 11 years) participated. Group 1 had 37 typically-developing (TD) hearing children (12 boys) whose ages at Time 1 ranged from 3.08 to 11.00 (mean: 7.30). Group 2 had 43 children with ASD (age range: 3.42 to 11.17; mean: 8.06; 34 boys) from specialist autism units located on the grounds of ordinary government-funded primary schools. Eligibility for these units had required extensive ASD screening by ASD-specialist clinicians who operated independently of our research team. Interviews, observations and tests had been administered and independently verified resulting in children with confirmed ASD diagnoses that fully met DSM criteria (APA, 2000/2013). The fact that the units were housed within ordinary primary schools meant that all with ASD in our sample had extensive daily opportunities for social interaction with ASD classmates and, at recess etc., with TD peers. Thus peer-based play and conversation was freely available and, on our informal observation, frequently undertaken.

Group 3 had 27 prelingually severely or profoundly deaf children (Time 1 mean age: 7.33; range 3.75 to 11.67; 16 boys). All were from hearing families where no one besides the child used sign language with as high a level of proficiency as a native speaker. These DoH children were pupils in specialist bilingual (Auslan/English) units attached to government-funded primary schools. At Time 1, all in Group 3 had at least a basic level of signing competence ("adequate for ordinary everyday communication" according to their teachers) and some also had some spoken language skills. However all these DoH children preferred signing to exclusive reliance on speech. Thus, with the aid of a professionally-qualified interpreter, they were tested bilingually in a signed-plus-spoken format, as detailed under Procedure below.

All three groups were recruited from government-funded preschools and primary schools in neighborhoods with similar socioeconomic (SES) catchments. Thus overall similarity in SES background across all groups was likely. All families had English as their sole or primary language and were of predominantly European-Australian ethnic background (approximately 86%) living in predominantly middle-class neighborhoods. By selecting the TD group to match the ASD and DoH groups by age, Time 1 ages of the three groups did not differ, F (2, 104) = 2.02, p =.138. At Time

2, the same was true, F (2, 104) = 2.72, p =.070. As reported later, the groups were also well-matched in language ability on standardized published language ability tests at Time 1.

Close group matching was also true of the interval between each child's initial and final test. Mean gaps (in years) between Times 1 and 2 were 1.23, 1.41, and 1.32 for TD, ASD and DoH groups, respectively, F (2, 104) = 2.11, p = .126. No child in any group had less than 10 months between tests, and the maximum in all three groups was the same: 25 months.

Tasks, Procedures and Scoring

ToM understanding. At Times 1 and 2, children individually took Peterson et al.'s (2012) 6-step ToM Scale (see Table 1) and, for comparison, a standard 3-item false belief battery (2 changed-location items from Baron-Cohen, Leslie and Frith (1985) and 1 misleading container from Wellman and Liu (2004)). As Table 1 shows, the 6-step ToM Scale had: (1) Diverse Desires (DD: the concept that different people can want different things), (2) Diverse Beliefs (DB: people's opinions about the same thing can differ), (3) Knowledge Access (KA: not seeing leads to ignorance), (4) False Belief (FB: standard misleading container task), (5) Hidden Emotion (HE: people can conceal their true feelings behind false facial expressions), and (6) Sarcasm (SARC: a message's intended meaning can differ from the literal meaning of its words). Past cross-sectional research has suggested that a ToM-based understanding of sarcasm is a particularly late-developing concept, difficult or bewildering even for some hearing and deaf adults (e.g., O'Reilly et al., 2014), Further, as noted earlier, cross-sectionally many TD children aged 8 to 11 years fail one or both of the 6-step scale's two final tasks (HE and SARC) suggesting its suitability for studying ToM growth in middle childhood. Methodologically, all six ToM Scale tasks shared similar formats, linguistic complexity and scoring. Conceptually, they were likewise alike in asking about a focal contrast between a mental state (what a protagonist wants, thinks, feels or intends communicating) and either external reality or someone else's thoughts. For both the ToM Scale and the false belief battery, we required correct responses to all control and test questions to pass any given task.

Summary scores. Several summary scores were possible: Each child's ToM Scale total was the total number of tasks passed (out of 6). The total false belief (TFB) score (0-3) was the sum of false belief tasks correct. There was also a total ToM (TotToM) composite score that summed the totals for the ToM Scale with the two changed-location false belief tasks. (Deliberately, the misleading container false belief task was only used once in this composite.) TotToM could (and did) range from 0 to 8 and its Cronbach alphas revealed sound internal consistency at Time 1 (0.77) and Time 2 (0.73) for the whole sample, with similar values for each subgroup.

Language ability. At Time 1, two separate standardized norm-referenced tests: (a) the Peabody Picture Vocabulary Test (PPVT: Dunn & Dunn, 1997) and (b) the 22-item syntax subscale of the Clinical Evaluation of Language Fundamentals (CELF-P: Wiig, Secord & Semel, 1992) were

used to estimate children's linguistic maturity at the start of the study. The PPVT uses picture-pointing responses to an age-graded set of orally-presented words. It is suitable for spoken English and has been used effectively in much prior research with children with typical development and ASD (e.g., Happé, 1995; Milligan, Astington & Dack, 2007). However it was unsuitable for our signing deaf children for several reasons, including (a) the prevalence of items with intuitively-obvious (or "iconic") Auslan signs (e.g., pointing at the elbow is both the correct PPVT response and the correct Auslan sign for the word "elbow"), and (b) the absence of discrete Auslan words (as per standard Auslan dictionaries) for many PPVT items. (This means these then can only be presented in sign if they are spelled out letter by letter). Such a format would greatly increase the difficulty of the test for deaf children—or hearing ones for that matter.

Thus, for all the deaf children and some children in each of the other groups, we used the 22-item syntax scale of the CELF-P (Wiig, Secord & Semmel, 1992) which has also been used effectively in prior ToM research with TD children (e.g., Ruffman, Slade, & Crowe, 2002) and is uniquely suitable for validly assessing linguistic maturity in Auslan (see Wellman & Peterson, 2013; Peterson et al., 2016). It assesses a broad range of developmentally-sequenced lexical, morphological, and syntactic concepts (including verb tense, relative clauses and embedded complement phrases) several of which involve syntax similar to the ToM test questions. For each group on each test the child's language measure was the raw total score (total items correct) not age-normed or transformed scores (like VMA). All deaf children (100%) had language data as did 37 of those with ASD (86%) and 65% of the TD group. (Missing data were due to essentially random factors--e.g., difficulties scheduling additional testing sessions late in the school year).

Our purpose in using the language tests was simply to rule out this variable as a possible confound rather than to precisely calibrate each child's linguistic maturity as a primary focus. Therefore, given our use of raw scores on both tests, we were able to provide a common metric for language ability across all participants by following the standardization procedures validated by Peterson et al. (2016). In brief, we first standardized raw score distributions for each test across the whole sample then assigned the standardized (*z*) score as the child's language ability measure. For children who took both tests (3 ASD and 2 TD) we conservatively took the lower standard score.

Although matching by language ability had not been a requirement for sample selection, it is useful that the groups (see Table 2 for means) did not differ significantly from one another in this respect, F (2, 86) = 2.49, p =.089. However, ASD children had a wider range of individual difference than the other groups, from almost 3 (2.96) standard deviations (SD) below to 1.55 SDs above the mean. This is consistent with contemporary diagnostic practice for ASD because deficits in *structural language* (i.e., slow vocabulary and/or syntactic growth) are no longer considered diagnostic whereas deficits in *pragmatic language* (social communication) still are (DSM-5: APA,

2013). Thus contemporary ASD populations typically include some with language delay and some without it, unlike formerly (e.g., under DSM-IV: APA, 2000) when autism was differentiated from Asperger Disorder via presence of structural language problems.

Recruitment and testing procedures. With schools' permission, we recruited children via an informative invitation letter sent home with the child. No child took part unless a parent supplied written informed consent for Time 2 as well as Time 1. Children gave their own verbal assent as a further precondition for participating. No payment or other incentives were promised to children or their families as inducements for participation but at the end of each session the child received a collectable paper sticker as a "thank-you" token.

All assessments were administered individually to each child in a quiet school area. For the deaf children all testing was fully bilingual. The main (hearing-speaking) experimenter was assisted by a highly professionally-qualified sign-language interpreter with native-like proficiency in Auslan. In a format familiar to the children in their everyday school routines, for each question or narrative segment, the main experimenter first spoke the relevant utterance with his lips clearly visible. The interpreter then immediately translated this spoken statement into the child's preferred signing modality. (This was either Australian Sign Language (Auslan) or Natural Sign System (NSS). The latter entails 'signing-in-English' via Auslan signs presented in spoken-English word order: Schembri & Johnston, 2007). Interpreters all had full professional accreditation at the "interpreter" level (formerly Level 3 of a 3-level scale), by the peak accreditation body in Australia, National Accreditation Authority for Translators and Interpreters (NAATI, 2011). All were likewise extensively experienced in working with deaf children in schools. Children with ASD and TD took the tests in spoken English only. With a few exceptions, data were collected between August 2009 and September 2016.

The project was ethically cleared by human research ethics committees of the university, the local government education authority, the schools' governing boards and all other relevant bodies. Procedures fully met international guidelines for ethical research with children (Graham, Powell & Taylor, 2015), including (a) written parental consent and child's verbal assent (see above), (b) continual monitoring of children's ongoing comfort and willingness to continue (had any sign of reluctance emerged, procedures would have been immediately discontinued), and (c) "debriefing" in the form of non-specific, non-evaluative positive feedback at the end of each session (e.g., "You did really well today/I have really enjoyed being able to talk with you today/ Your help with our project is very important to us").

Results

Overview

Several preliminary analyses were needed before addressing our primary substantive questions. Thus results are considered in four sections. Section 1 begins with an analysis of the longitudinal validity of the 6-step ToM Scale, first for TD children and then for those with deafness and ASD. Second, we provide an omnibus analysis including all three diagnostic groups to compare ToM performance at each testing time between children with ASD, deafness, and typical development and also the extent of longitudinal gain by each group. This enables us to ask: Do most individuals in all these groups continue to develop longitudinally? Or, for at least some groups, are longitudinal gains either non-existent or limited to a very small minority of individuals? Following this we examined substantive questions within each group including the above-noted questions of theoretical relevance concerning; (a) comparative longitudinal changes for younger versus older children and (b) longitudinal stability of within-group individual differences. We address these questions first for TD children then for ASD and DoH groups. Finally, fourth, we use hierarchical multiple regression to see what variables (including, but not limited to, deafness or ASD status) may help to determine whether children (a) will or will not make longitudinal ToM progress and (b) the extent of that progress.

Longitudinal Evidence of 6-Step ToM Scale Progressions

Figure 1 provides a graphical depiction of the results, useful as background for all our statistical analyses.

TD children. Table 1 shows pass rates for each ToM Scale task for the TD children at Times 1 and 2. Patterns of task success and failure at each testing time conform to the predicted order of difficulty (i.e., DD> DB> KA> FB> HE> SARC) obtained from past cross-sectional research. In fact, at Time 1, 97% of the 37 TD children conformed perfectly to the entire scale sequence across all its six steps. At Time 2, 92% continued to do so. As Figure 1 shows, the trajectory of overall task success over time was likewise consistent with an orderly increase in total tasks passed over development.

Guttman scaling statistics (Green, 1956) assess a sample's observed patterns of success and failure against the perfect sequences that would emerge if each child passed or failed each ToM Scale task in the precise scale pattern such that no more advanced task would be passed once an earlier one was failed. Green's (1956) coefficient of reproducibility (Rep) statistically evaluates how closely an observed set of data match this ideal of perfect scale conformity and is significant at values of .90 or higher. Conformity to the DD> DB> KA> FB> HE> SARC sequence was highly significant in both cross-sectional samples of this cohort sequential design for TD children aged 3 to 13. At Time 1 Rep was .995 and at Time 2 it remained high at .986.

For children who are perfectly scale consistent at Time 1 (97% of these TD children) it is easy and revealing to examine their longitudinal progressions in more detail. The key question is

how many proceed along the scale in the prescribed order versus how many deviate by either going backwards through the scale or skipping a step to pass a harder step out of order after failing a previous one. Children who scored identically at the two times count as proceeding in order since they evidenced perfectly scale-consistent patterns that neither went backwards nor skipped over incorrect items (see Wellman et al. 2011). For TD children, 84% proceeded in perfect order from Time 1 to Time 2. Thus, clear longitudinal consistency in scale sequencing emerged, identical to the sequential patterns indirectly manifest in the present and prior cross-sectional data for TD groups.

A consideration for examining ToM development during early and middle childhood concerns ceiling and floor effects. At Time 1, despite ages as low as 3 years 1 month, no TD child scored at floor (the lowest scale score was 2). Ceiling effects at Time 1 were likewise minimal despite ages ranging up to 11 years. Only 9 of the 37 TD children (24%) scored perfectly by passing all six of the ToM Scale tasks at Time 1. Thus there was room for almost all of them to make longitudinal improvements on the ToM Scale (and results showed that they did so).

In contrast, TFB was methodologically weaker through its high proportion of Time 1 ceiling effects among the TD children who were age-matched to the ASD and DoH groups. In fact, when we checked for ceiling effects in our data, the vast majority (93%) of the younger subgroup of TD children (n = 14) who were aged 3 through 6 years showed no ceiling effects. However most (96%) of those aged 7 to 11 (n = 23) were at Time 1 TFB ceiling. Thus TFB was not used separately for further analyses.

DoH children. Past cross-sectional research (e.g., Peterson et al., 2012; Peterson, Slaughter & Wellman, 2017) shows that DoH children's 6-step ToM Scale sequence matches TD children's (namely: DD> DB> KA> FB> HE> SARC). Table 1 confirms this sequence for the current DoH group at both testing times. Individual DoH children's patterns of responding to the six tasks also mostly conformed perfectly to the predicted sequence. At Time 1, 17 of the 27 DoH children (63%) displayed perfect scale conformity over all six tasks. At Time 2, 82% did so. Guttman scaling methods again confirmed the statistical reliability of this perfect scale conformity. At Time 1, the DoH group's Rep coefficient was .93 and at Time 2, their Rep was .96, both statistically significant.

Paralleling the TD analyses, we examined longitudinal scale progressions over time by looking at DoH children who were perfectly scale consistent at Time 1. All 17 (100%) proceeded longitudinally through the sequential steps of the scale in the prescribed task order without going backward or skipping over a failed task to pass one later in the sequence. Figure 1 illustrates their consistently upward longitudinal ToM trajectory, similar to TD children's.

Children with ASD. Past cross-sectional research on children with autism (Peterson et al., 2005; 2012; 2017) has suggested an alternative scale sequence for this group (DD> DB> KA> **HE>**

FB> SARC) where children with ASD often fail false belief while passing HE. Response patterns for our ASD group at each testing time (see Table 1) conformed to this expected ASD-specific ordering. Combining both time points together, there were 17 ASD children who passed either HE or FB but not both. Of these, a strong majority (65%) passed HE while failing FB compared with only 35% passing FB but not HE. By contrast, all 17 TD children (100%) who passed only one of these two tasks passed FB only. This TD-versus-ASD difference in relative success rates on FB versus HE (for children passing only one or other task) was statistically significant, *Chi square* (1) = 16.26, N = 34, $p < .001$.

Moreover, 79% of the ASD children at Time 1, and 91% at Time 2, conformed perfectly to their alternative (HE> FB) sequence across all its six scale steps. Guttman scaling thus revealed a Rep coefficient for the ASD group at Time 1 of .97 and at Time 2 of .98, both statistically significant. Likewise, out of the 34 ASD children who were perfectly scale-consistent at Time 1, 32 (94%) proceeded longitudinally through the scale in exactly the prescribed sequence. In short, the 6-step scale captured ASD children's individual patterns of longitudinal progress, just as for the other two groups, albeit conforming to their alternative sequence. As Figure 1 shows, total ToM scores for ASD children also displayed an overall upward longitudinal trajectory, though over a broader age range than for the TD group. Just like the other groups', ASD children's ToM Scale totals and TotToM scores were suitably free at Time 1 of both ceiling effects (2%) and floor effects (5%) throughout the age range we studied.

Using the 6-Step ToM Scale to Explore Children's Longitudinal ToM Development

Longitudinal validation of the 6-step scale for all three diagnostic groups in our sample has importance in its own right, but more importantly allows addressing our substantive questions.

Comparisons among groups at Times 1 and 2. Table 2 shows the three groups' mean ToM Scale scores and their Total ToM composite at Times 1 and 2 along with background variables (age, language ability, and gap (in months) between Times 1 and 2). Groups did not differ significantly in any of these background variables (see Method). We therefore used one-way ANOVAs, with Newman-Keuls post-hocs, to compare the groups' ToM performance at each testing time. Groups differed significantly on the 6-step ToM Scale at Time 1, F (2, 104) = 31.62, $p < .001$, and Time 2, F (2, 104) = 18.22, $p < .001$. TD children significantly outperformed both atypical groups at both times, while ASD and DoH groups scored equivalently (see Figure 1).

TotToM also differed by group at Time 1, F (2, 104) = 35.81, $p < .001$, and Time 2, F (2, 104) = 22.43, $p < .001$, with TD children again significantly outperforming ASD and DoH peers who scored equivalently. On the ToM Scale, after excluding the 10 Time 1 ceiling performers (9 TD; 1 ASD) there was no group difference in amount of gain between Times 1 and 2, F (2, 94) = 1.20, $p = .240$, despite lower baseline starting levels for those with deafness or ASD. Further, at

Time 2, the ASD and DoH groups remained substantially behind their TD peers on the ToM Scale, confirming that atypical children's ToM delays persist through childhood and apply broadly to all six ToM Scale concepts rather than being confined to just the preschool period or just false belief. Consistent past cross-sectional research, ASD and DoH children showed no sign of catching up to TD peers (between-group catch up as opposed to within-group catch up) nor even of narrowing the gap.

Developmental gain through middle childhood. Given these favorable ToM Scale properties and sample-wide results, examining cross-sequential development through middle childhood for each group separately has theoretical interest. For TD children, ToM Scale scores at Time 2 (Table 2) significantly exceeded Time 1's, t (36) = 3.61, p = .001. This was also true for ASD children, t (42) = 3.61, p = .001, and DoH children, t (26) = 6.19, p< .001. Thus significant longitudinal progress, in a reliably sequenced developmental progression, arose in each separate group via mastery of fresh ToM concepts rather than just increasingly accurate performance on a single concept like false belief. To look specifically at longitudinal ToM Scale development *during* middle childhood for TD children, we focused on the 23 aged 7 to 11 (mean = 8.44; exact range: 86 to 132 months) who were older than the conventional age of false belief mastery. A matched-pair t test showed a significant increase in ToM Scale scores between Times 1 and 2, t (22) = 2.86, p = .009, and this remained true with the 9 Time 1 ceiling performers eliminated, t (14) = 3.16, p = .007. Thus, even the oldest TD cohort mastered fresh ToM concepts longitudinally during middle childhood. Comparatively, the 12 DoH children aged above 7 also had significantly higher ToM Scale scores at Time 2 (mean = 4.08) than Time 1 (mean = 2.75: t (14) = 5.93, p < .001), as did the 29 with ASD aged 7 and older (Time 1 mean: 3.48; Time 2 mean: 4.14; t (28) = 3.62 p = .001).

Children's relative position in their group over time. Exploring the stability of TD children's individual differences over time (i.e., their relative position in being ahead of, equal to or behind other TD children their age), we excluded the 9 with ceiling Time 1 ToM Scale scores and subdivided the remaining 28 into those initially scoring higher (5 or 6 steps passed at Time 1) versus lower (0 to 4 steps). Following Devine et al. (2016), three patterns are possible: (a) "within-group acceleration": children with *higher* ToM scores initially make more gain (gain score would be higher), (b) "within-group catch-up": children with *lower* scores initially make more gain (gain score would be higher) and (c) "within-group even gains": children with higher or lower scores initially both make similar gains (gain scores are not different). In fact, TD children's results supported the latter. Lower-scoring (n=13) and higher-scoring (n=15) subgroups made similar gains (means = .46 and .73, respectively; t (26) < 1.00, p = .340), consistent with an "even-

gains" model and with Bornstein et al.'s (2014) findings regarding the stability of individual differences in TD children's language development across a similar age range.

Similarly, in the ASD group (after excluding the one ceiling performer), the 19 children with lower Time 1 scores (0 to 3 steps) gained a mean gain of 1.00 steps versus a mean gain of .52 steps for those (n= 23) with higher Time 1 scores (4 to 6 steps), a nonsignificant difference, t (40) = 1.64, p = .109. In the DoH group, the Time 1 lower-scorers (n =19: 0 to 2 steps) gained a mean of 1.00 steps compared with a mean of .88 for the 8 who were higher-scoring (3+ steps passed at Time 1). Again the difference was not significant, t (25) = .47, p = .254.

Given these similarities, we increased the statistical power of the comparison by combining the TD, ASD and DoH groups for a 2 (high/low Time 1 ToM) x 3 (group: TD, ASD or DoH) ANOVA on gain scores for the 97 children who were below ceiling at Time 1. Mean gains for low scorers (n = 41) and high scorers (n = 56) were .98 steps and .76 steps, respectively. There was no significant main effect of group, F (2, 91) < 1, p = .590, nor of Time 1 score category, F (2, 91) < 1, p = .335, nor of the interaction, F (2, 91) < 1, p = .648. In this analysis as well, the even-gains conclusion applies generally to all three groups.

Within-group individual differences in the possibility for progress. How widespread are developmental gains across individual members of each group? Do groups differ in this respect? One hypothesis, for children with autism (e.g., Holroyd & Baron-Cohen, 1993), is that only a small minority of exceptional individuals are capable of making any longitudinal ToM progress at all. An alternative hypothesis, consistent with past evidence for environmental influences (e.g., social exchange of mentalistic conversation) on ToM over time not only for TD children but also for those with ASD and deafness (Siegal & Peterson, 2008), is that development continues for all or most individuals across a 1-to 2-year interval despite atypical groups starting at a lower level initially. To examine this, we categorized each child's longitudinal pattern based on numbers of ToM Scale steps passed at Times 1 and 2. The "gain" category reflected a higher Time 2 than Time 1 total. "No change" reflected exactly the same score (identical numbers of steps passed) at Times 1 and 2 and "decline" reflected a lower total score at Time 2 than Time 1. Then, to compare the groups, we again set aside the 10 Time 1 ceiling performers (9 TD; 1 ASD). The remaining 97 performed as Table 3 shows. On the ToM Scale, at least half the children in each group made longitudinal progress, mastering at least one additional scale step beyond where they had started at Time 1. The rest essentially displayed no longitudinal change. Only 3 children (1 TD, 2 ASD) regressed developmentally to score lower at Time 2 than Time 1. Combining non-changers with decliners, numbers of children progressing versus not progressing did not differ by group either on the ToM Scale, *Chi square* (2) = 2.46, N = 97, p = .292, or on TotToM, *Chi square* (2) = 1.68, N = 97, p =

.431. This contradicts the notion that if longitudinal ToM gain occurs at all in connection with ASD (or DoH) it is confined to only a tiny minority of individuals.

Variables Predicting Longitudinal ToM Progress

To comprehensively explore predictors of individual differences in children's longitudinal ToM gains we used a hierarchical multiple regression analysis with Time 2 TotToM as the dependent variable. Entry at Step 1 of Time 1 age, Time 1 language ability and gap (the interval between the child's initial and final test) as control variables produced a significant equation, F (3, 83) = 17.27, $p < .001$. Beta weights showed that both age (*beta* = .20, $p = .037$) and language ability (*beta* = .50, $p < .001$) were significant predictors but the gap between testings was not. Next, at Step 2, the Time 1 TotToM score was entered as a predictor. There was large increment with this addition to the model, F (change) = 95.52, $p < .001$. The overall equation was also significant at this step, F (4, 82) = 51.58, $p < .001$. Beta weights at Step 2 indicated that the only significant predictor of Time 2 TotToM scores in this model was Time 1 TotToM (*beta* = .74, $p < .001$). Age (beta =.05, $p = .436$) and language ability (*beta* =.13, $p = .091$) fell to non-significance. This shows that children with better ToM understanding at Time 1 were the ones who scored highest at Time 2 irrespective of their age, language ability or the interval between their tests.

Of course this overall result could conceivably vary, say for ASD children, or DoH children, in some telltale way. Results at the final step of the hierarchical regression model (see Table 4) suggest otherwise. For this third step, ASD status and deafness status were entered as predictors, dummy-coded (1 for "present" or 0 for "absent") as recommended by Tabachnik and Fidell (2001). No significant increment arose at this step, F (change) = 1.17, $p = .314$, although the full equation remained significant, F (6, 80) = 34.92, $p < .001$. The only significant beta weight in the final model (see Table 4) was Time 1 TotToM, indicating neither being deaf nor having ASD changed the pattern observed at Step 2.

Discussion

Methodologically, our cross-sequential data supplied impressive longitudinal (as well as cross-sectional) validation of the 6-step ToM Scale as a sensitive measure of developmental change. Longitudinal progress through middle childhood for both typically and atypically developing children was reliably scale-consistent. This methodological conclusion, and these findings, are strengthened because they hold across three disparate groups varying both in their sequences of ToM mastery (e.g., TD and DoH versus ASD children), their starting points (e.g., ASD and DoH groups both below TD children) and their final end points after 1 to 2 years of development (ASD and DoH children below TD children).

More substantively, several key findings emerged. As a cornerstone of social intelligence and satisfying social interaction, ToM develops rapidly not only during preschool but also in middle

childhood. This is particularly true for ToM Scale tasks beyond false belief (with its ceiling limitations for older TD children). ToM development through the "uncharted waters of middle childhood" (Hughes, 2016, p. 4) is evident not just for TD children but even more so in the context of ToM delay. Because our longitudinal evidence confirms that the 6-step ToM Scale is an informative yardstick for measuring children's ToM progress well beyond preschool, our data can address several theoretical questions about ToM acquisition. In the introduction, we outlined three contrasting hypotheses about ToM growth during early as compared with later childhood. One possibility was that rapid early ToM gain could taper off at the end of preschool, once the false belief milestone was mastered. In relation to our scale, this might manifest itself as one year of aging achieving more ToM steps before age 6 than an equivalent one year at later (post-preschool) ages. Alternatively, there could be acceleration of ToM growth after preschool as the more complex social demands of school culture and classroom peer groups pose new challenges to stimulate ToM reasoning. Or, conceivably ToM progress could continue steadily at the same rate throughout childhood. This last was essentially what we found (see Figure 1). Thus, older TD children evidenced orderly longitudinal gain on the ToM Scale just as preschool children did.

Furthermore, even though ASD and DoH children remained well behind their TD peers at the study's start and at its end, gains arose similarly steadily for both these atypical groups throughout childhood. For atypical groups, in particular, it was conceivable in advance of our data that development might cease for some or many. That is, progress up to and beyond false belief would become increasingly difficult or might never be achieved by most such children (Ozonoff & McEvoy, 1994). Yet, contrary to this theory, we found that later ToM Scale concepts positioned higher up the scale (see Table 1) were just as attainable for all three groups as the initial "preschool" steps that most TD children master by age 6. Our data were able to confirm this possibility because of their Guttman scaling properties where each successive scale step represents a discrete conceptual advance upon the previous step. Thus the 6-step ToM Scale is more than a simple battery of cognate tasks (as TFB is). Instead, it validly measures progressively more advanced ToM milestones that occur in a reliable developmental sequence. This was confirmed both cross-sectionally and longitudinally in our data for both TD and ToM-delayed groups.

Our focus on individual differences addressed (a) whether longitudinal progress was widespread among most children in a group versus limited to a few exceptional individuals and (b) whether initial differences among individuals are predictive of later individual differences. Results showed firstly that, for both TD and ToM-delayed children, developmental gain was widespread. A majority in all groups who had room to progress actually did so. Furthermore, progress over the mean of 1.5 years was substantial, amounting to a gain of at least one full ToM Scale step for over half the children in each group. In terms of within-group relative standing, our data favor an "even

gains" (or "stable individual differences": Bornstein et al., 2014) hypothesis, rather than, for example, a within-group "catch-up" mechanism where children with a slow initial ToM trajectory might come to equal or overtake their peers later on. Progress by individuals who started the study ahead of others in their diagnostic group was neither faster nor slower than for those who began relatively behind. Of course these conclusions need to be qualified by limitations of our modest sample sizes (especially for TD children aged 3 to 6 years) and the possibility that individual differences at ages outside of the range we examined might show a different pattern. Further longitudinal research with the 6-step ToM Scale could helpfully explore these possibilities more fully.

Results of the hierarchical multiple regression analysis showed that the best predictor of children's final composite ToM score at Time 2 was their initial level of ToM understanding. Over and above age, language ability and disability status, Time 1 ToM understanding significantly predicted subsequent ToM 1 to 2 years later. Many to most children in all groups made gains over the longitudinal period, appearing to build upon their earlier ToM achievements to progress steadily forward through subsequent ToM milestones. This is graphically illustrated in Figure 1.

Atypical ToM Development

Our inclusion of children with autism or deafness representing preschool age through late childhood provides more than a methodological confirmation of the above conclusions across a wide swath of development by individuals of widely varying levels of initial ToM abilty. It also allows us to address other theoretical questions about typical and atypical ToM development.

DoH children. Our inclusion of DoH children seems to us especially informative. Because deaf children's atypicality consists merely of peripheral auditory deficits, they lack the central neurological abnormalities characterizing children with autism. Thus their ToM Scale progressions arguably more clearly and precisely reflect the influences of cultural and social experiences (including conversation, play and interpersonal interaction) without the complication of neuro-cognitive atypicality. Given the known impact of such language-saturated social experiences for TD children's ToM development (e.g., Harris, 2005), then research on DoH children who are initially limited in these experiences can critically address how social-interactive-linguistic factors can influence a cascade of understandings about minds in interconnected ways. This approach to looking at deaf children's data is enhanced in our study because these DoH children demonstrated a typical progression of gradually unfolding ToM conceptual steps, albeit with delay. Admittedly, we did not collect social-conversational data for any of our groups, so this interpretation rests, as is common practice, on known differences in the conversational experiences of DoH children as a group, differences that have been validated repeatedly in other research that has examined parent-child conversation (e.g., Harris, 2006; Slaughter & Peterson, 2011).

ASD children. Much cross-sectional research shows that children with autism differ informatively from TD groups in the timing of their ToM mastery (see: Baron-Cohen, 1995; Happé, 1995; Siegal & Peterson, 2008 for reviews). Our data confirm this but also extend such conclusions longitudinally. Over the range 3 to 12 years, our cross-sequential data showed children with ASD to be significantly delayed behind TD children on a progressive set of ToM understandings, as were DoH children. Furthermore, both ASD and DoH groups remained well behind their age-matched TD peers at the end of the study. Yet ASD and DoH groups both continued to develop longitudinally at an equivalent rate throughout childhood (see Figure 1).

However, these similar delays and rates of gain for ASD and DoH children could be for different underlying theoretical reasons. This possibility might connect with ToM sequence differences (notwithstanding timetable similarity) between these groups. Indeed the variation we observed in the sequential ordering of ASD children's ToM Scale steps is instructive. Confirming past sparse and purely cross-sectional findings, our cross-sequential longitudinal data show an atypical sequential ordering in ASD that runs counter not only to the ordering for TD children but importantly also contrary to DoH children's. For the ASD group, understanding of false belief emerged one developmental step later than hidden emotion. The DoH versus ASD difference is especially important because both these groups are equally delayed. While suggestive of an important cognitive–developmental difference between those with autism versus deafness, possible underlying explanations are varied. The unique neurological atypicalities of autism cannot be ruled out. Plausibly, however, sequence differences might instead reflect differences between DoH and ASD children's environments and social experiences (e.g., ASD children's greater victimization by peers' teasing: Peterson et al., 2005). Clearly, the best interpretation of these intriguing findings awaits further investigation.

One conclusion we can advance with certainty is that most ASD children, like their TD and DoH peers, do continue to make substantial longitudinal ToM progress during the school years. Strikingly, some past longitudinal studies of autistic people's false belief understanding (e.g., Holroyd & Baron-Cohen, 1993; Ozonoff & McEvoy, 1994) found a lack of any consistent ToM progress by a large majority in this diagnostic group even at very advanced child and adolescent ages and even across 3- to 8-year longitudinal time spans. Exclusive reliance on the false belief test in these past studies may have been partly responsible. Indeed, our data were very different, revealing gains of at least one full step on the ToM Scale by slightly more than half of these children over an average of just 1.5 years. In our data, TFB (total false belief) proved less revealing than the ToM Scale of longitudinal progress. For example, at Time 1, 57% of our ASD group's responses were at floor on TFB and 4% were at ceiling. Yet, on the ToM Scale, this was true of

only 4% and 2%, recommending the scale as a more sensitive index than just false belief for future longitudinal research with ASD children. (Nevertheless, even on TFB our ASD children scored significantly higher at Time 2 than they had at Time 1, z (Wilcoxon) = 3.45, p = .001).

Overall, longitudinal use of the 6-step ToM Scale supplied new insights for ASD, DoH and TD children. These insights are suggestively optimistic, highlighting widespread developmental possibilities even in the context of severe ToM delay. Contrary to some earlier research, we found many to most deaf and autistic children to be capable of genuine ToM growth, even at ages well beyond preschool.

These and other novel results of our study provoke numerous further questions about typical and atypical longitudinal ToM development that await future research. For example, our finding that even TD children aged 7 to 12 years were continuing to make statistically significant ToM progress is at odds with the notion of an experience-dependent sensitive period confined to the preschool years (e.g., Siegal & Varley, 2002). However, conceivably the sensitive period's upper age boundary could extend beyond age 12. To explore this and other provocative results of our study in greater depth, future studies are clearly needed.

These should examine TD, ASD and DoH groups' ToM Scale development over an even wider age range than in our study, and over a longer longitudinal period, perhaps beginning at a later mean age. Our longitudinal interval (1 to 2 years) was long relative to many past longitudinal studies of TD and DoH children. Yet two previous longitudinal ASD studies (Holroyd & Baron-Cohen, 1993; Ozonoff & McEvoy, 1994) used substantially older samples and longer intervals (8 and 3 years, respectively). Had any evidence of longitudinal gain emerged in either of these studies, interpretation of our results could have been affected, especially had our ASD sample failed to progress. Quite the contrary, however, our ASD group *did* progress over just 1.5 years (52% gained one ToM Scale step or more) whereas both these earlier studies showed *no* overall progress over intervals at least twice as long. Nevertheless future longitudinal tracking of ASD and DoH individuals over longer time frames than in our study is undeniably needed. Hopefully such studies will include older groups and will use the 6-step ToM Scale, given its current longitudinal validation. Especially important will be to explore whether longitudinal ToM Scale gains (a) continue in teen and adult ASD and DoH groups and (b) remain as widespread amongst individuals as we found, or (perhaps) become even more so.

References

American Psychiatric Association (APA). (2000). Diagnostic and statistical manual of mental disorders (4th ed.). doi:10.1176/appi.books.9780890423349.

American Psychiatric Association (APA). (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Association. doi:10.1176/2013-14907-000.

Bagwell, C. L., Newcomb, A. F., & Bukowski, W. M. (1998). Preadolescent friendship and peer rejection as predictors of adult adjustment. *Child Development, 69*, 140-153. doi: 10.1111/j.1467-8624.1998.tb06139.x

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

Baron-Cohen, S. (2001). Theory of mind development in normal development and autism. *Prisme, 34*, 174-183.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*, 37-46. doi:10.1016/0010-277(85)900228

Bornstein, M., Hahn, C., Putnick, D. & Suwalksky, J. (2014). Studies of core language skills from early to late childhood. *Child Development, 85*, 1346-1356. doi: 10.1111/cdev12192

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S. &Singh, S. (2005). Synchrony in the onset of mental-state reasoning. *Psychological Science, 16*(5), 378-384. doi:10.1111/j.0956-7976.2005.01544.x

Capage, L. & Watson, A. (2001). Individual differences in theory of mind, aggressive behavior and social skills. *Early Education & Development, 12*, 613-628.

Caputi, M., Lecce, S., Pagnin, A., & Banerjee, R. (2012). Longitudinal effects of theory of mind on later peer relations: The role of prosocial behavior. *Developmental Psychology, 48*, 257-270. doi:10.1037/a0025402

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

Devine, R. T., & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child Development, 84*, 989-1003. doi:10.1111/cdev.12017

Devine, R. T., White, N., Ensor, R., & Hughes, C. (2016). Theory of mind in middle childhood. *Developmental Psychology, 52*(5), 758-771. doi:10.1037/dev0000105

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-mind training causes honest young children to lie. *Psychological Science, 26*, 1812-1821. doi: 10.1177/0956797615604628

Dunn, J. (1995). Children as psychologists: The later correlates of individual differences in understanding of emotions and other minds. *Cognition & Emotion, 9,* 187-201. doi: 10.1080/02699939508409008

Dunn. L.M. & Dunn, L.M. (1997). Peabody Picture Vocabulary Test (Third Edition: PPVT-III). Circle Pines, MN: American Guidance Service.

Fink, E., Begeer, S., Peterson, C. C., Slaughter, V., & Rosnay, M. (2014). Friendlessness and theory of mind: A prospective longitudinal study. *British Journal of Developmental Psychology*. doi: 10.1111/bjdp.12080

Flavell, J. H. (2004). Theory of mind: Retrospect and prospect. *Merrilll-Palmer Quarterly, 50,* 274-290. doi: 10.1353/mpq.2004.0018

Graham, A., Powell, M. & Taylor, N. (2015). Ethical research involving children. *Family Matters, 96,* 23-28.

Green, B. (1956). A method of scalogram analysis using summary statistics. *Psychometrika, 21*(1), 79-88. doi:10.1007/bf02289088

Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66,* 843-855. doi: 10.2307/1131954

Harris, P. L. (2005). Conversation, pretense and theory of mind. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 70-83). New York: Oxford University Press.

Heyes, C. (2014). False belief in infancy: Afresh look. *Developmental Science, 17,* 647-659. doi: 10.1111/desc.12148

Holroyd, S., & Baron-Cohen, S. (1993). Brief report: How far can people with autism go in developing a theory of mind? *Journal of Autism and Developmental Disorders, 23*(2), 379-385. doi:10.1007/BF01046226

Hughes, C. (2016). Theory of mind grows up. *Journal of Experimental Child Psychology, 149,* 1-5. doi:10.1016/j.jecp.2016.01.017

Leslie, A. & Thaiss, L. (1992). Domain specificity and conceptual development: Neuropsychological evidence from autism. *Cognition, 4,* 225-251. doi: 10.1016/0010-0277(92)90013-8

Meltzoff, A. (1995). Understanding the intentions of others. *Developmental Psychology, 31,* 838-850. doi: 10.1037/0012-1649.31.5.838

Meristo, M., Falkman, K., Hjelmquist, E., Tedoldi, M., Surian, L., & Siegal, M. (2007). Language access and theory of mind reasoning. *Developmental Psychology, 43,* 1156-1169. doi:10.1037/0012-1649.43.5.1156

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78,* 622-646.doi: 2007.01018x.1111/j.1467-86

Mundy, P., Sullivan, L. & Mastergeorge, A. (2009). A distributed systems model of joint attention, social cognition and autism. *Autism Research, 2,* 2-21. doi: 10.1002/aur.61

NAATI (National Accreditation Authority for Translators and Interpreters) (2011). Accreditation standards for Auslan interpreters. Retrieved 22/09/2012 from www.naati.com.au

Newport, E. (1991). Contrasting concepts of the critical period for language. In S. Carey & R. Gelman (Eds.). *The epigenesis of mind* (pp. 111-130). Hillsdale, NJ: Erlbaum.

O'Reilly, J. & Peterson, C.C. (2014). Theory of mind at home. *Early Child Development & Care, 184,* 1934-1947. doi:10.1080/03004430.2014.894034

O'Reilly, K., Peterson, C. C., & Wellman, H. M. (2014). Sarcasm and advanced theory of mind understanding in children and adults with prelingual deafness. *Developmental Psychology, 50,* 1862-1877. doi:10.1037/a00366

Osterhaus, C., Koerber, S. & Sodian, B. (2016). Scaling advanced theory of mind tasks. *Child Development, 87,* 1971-1991. doi:10.1111/cdev12566

Ozonoff, S. & McEvoy (1994). A longitudinal study of executive function and theory of mind in autism. *Development & Psychopathology, 6,* 415-431. doi:10.1017/S0954579400006027

Perner, J., & Wimmer, H. (1985). 'John thinks that Mary thinks that...': Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology, 39,* 437-471. doi:10.1016/0022-0965(85)90051-7

Peterson, C. C. (2009). Development of social-cognitive and communication skills in children born deaf. *Scandinavian Journal of Psychology, 50,* 475-483. doi: 10.1111/j.1467-9450.2009.00750.x

Peterson, C. C., O'Reilly, K., & Wellman, H. M. (2016). Deaf and hearing children's development of theory of mind, peer popularity, and leadership during middle childhood. *Journal of Experimental Child Psychology, 149,* 146-158. doi:/10.1016/j.jecp.2015.11.008

Peterson, C. C., & Siegal, M. (1999). Representing inner worlds: Theory of mind in autistic, deaf, and normal hearing children. *Psychological Science, 10,* 126-129. doi: 10.1111/1467-9280.00119

Peterson, C. C., Slaughter, V., Moore, C. & Wellman, H.M. (2106). Peer social skills and theory-of-mind development in children with autism, deafness or typical development. *Developmental Psychology, 52,* 46-57. doi:10.1037/a0039833

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. doi: 10.1111/j.1467-8624.2005.00859.x

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory of mind scales for typically developing children, and those with deafness, autism, or Asperger Syndrome. *Child Development, 83*, 469-485. doi: 10.1111/j.1467-8624.2011.01728.x

Pons, F., Harris, P. L., & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European Journal of Developmental Psychology, 1*(2), 127-152. doi:10.1080/17405620344000022

Poulin-Dubois, D. & Yott, J. (2017). Probing the depths of infants' theory of mind: Disunity in performance across paradigms. *Developmental Science,* 1-12. doi:10.1111/doc12600

Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science, 20*(7), 805-812. doi:10.1111/j.1467-9280.2009.02377.x

Razza, R. A., & Blair, C. (2009). Associations among false-belief understanding, executive function, and social competence: A longitudinal analysis. *Journal of Applied Developmental Psychology, 30*(3), 332-343. doi:http://dx.doi.org/10.1016/j.appdev.2008.12.020

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. doi: 10.1111/1467-8624.00435

Schembri, A., & Johnston, T. (2007). Sociolinguistic variation in the use of fingerspelling in Australian Sign Language: A pilot study. *Sign Language Studies, 7*(3), 319-347. doi:10.1353/sls.2007.0019

Schaie, K. W. (1972). Limitations on the generalizability of growth curves of intelligence: A reanalysis of some data from the Harvard Growth Study. *Human Development, 15*(3), 141-152. doi:10.1159/000271238

Schick, B., deVilliers, P., deVilliers, J. & Hoffmeister, R. (2007). Language and theory of mind: A study of deaf children. *Child Development, 78*, 376- 396. doi: 10.1111/j.1467-8624.2007.01004

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

Siegal, M., & Varley, R. (2002). Neural systems involved in 'theory of mind'. *Nature Reviews Neuroscience, 3,* 463-471. doi:10.1038/nrn844

Slaughter, V., Imuta, K., Peterson, C. C., & Henry, J. D. (2015). Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development, 86*(4), 1159-1174. doi:10.1111/cdev.12372

Slaughter, V. & Peterson, C. C. (2011). How conversational input shapes theory of mind development. In M. Siegal & L. Surian (Eds.) *Access to language and cognitive development* (pp. 3-22). New York: Oxford University Press doi: 10.1093/acprof: oso/9780199592722.003.0001

Slaughter, V., Peterson, C. C., & Moore, C. (2013). I can talk you into it: Theory of mind and persuasion behavior in young children. *Developmental Psychology, 49*(2), 227-231. doi:10.1037/a0028280

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology, 30*(3), 395-402. doi:10.1037/0012-1649.30.3.395

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics.* Boston, MA: Allyn and Bacon.

Tager-Flusberg, H. & Sullivan, K. (1994). A second look at second-order belief attribution in autism. *Journal of Autism & Developmental Disorders, 24*, 577-585. doi:10.1007/BF02172139

Tierney, A. & Nelson, C. (2009). Brain development and the role of experience in the early years. *Zero to Three, 30*, 9-13. PMCID: PMC3722610.

Vaccari, C. & Marschark, M. (1997). Communication between parents and deaf children: Implications for social-emotional development. *Journal of Child Psychology and Psychiatry, 38,* 793-801. doi:10.1111/j.149-7610-1997.tb01597.x

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. New York: Oxford University Press.

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. doi:10.1111/j.1467-8624.00304

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory of mind scale: Longitudinal perspectives. *Child Development, 82*(3), 780-792. doi: 10.1111/j.1467-8624.2011.01583.x

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*, 523-541. doi:10.1111/j.1467-8624.2004.00691.x

Wellman, H. M., & Peterson, C. C. (2013). Deafness, thought bubbles, and theory-of-mind development. *Developmental Psychology, 49*(12), 2357-2367. doi:10.1037/a0032419

Wellman, H.M. & Woolley, J. (1990). From simple desires to ordinary beliefs. *Cognition, 35*, 245 - 275. doi:10.1016/0010-0277(90)9024-E

Wiig, E., Secord, W., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals—Preschool*. San Antonio, TX: Psychological Corporation, Harcourt Brace.

Woolfe, T. Want, S. & Siegal, M. (2002). Signposts to development: Theory of mind in deaf children. *Child Development, 73*, 768-778. doi10.1111/1467-8624.00437

Table 1. ToM Scale tasks and numbers (and percentages) of children in each group passing them at Times 1 and 2

Task	Diverse Desires	Diverse Beliefs	Knowledge Access	False Belief	Hidden Emotion	Sarcasm
ToM Concept	Different people want different things	People's (possibly true) beliefs can differ	Seeing leads to knowing; not seeing to ignorance	People can believe things that are not true	People can conceal their true feelings behind false expressions	People can mean the opposite of what they say
Group 1: TD (n = 37)						
Time 1:	37 (100%)	36 (97%)	35 (95%)	30 (81%)	22 (59%)	9 (24%)
Time 2:	37 (100%)	36 (97%)	36 (97%)	34 (92%)	24 (65%)	18 (49%)
Group 2: ASD (n =43)						
Time 1:	41 (95%)	36 (84%)	27 (63%)	9 (21%)	11 (26%)	3 (4%)
Time 2:	42 (98%)	41 (95%)	33 (77%)	17 (40%)	20 (46%)	5 (12%)
Group 3: DoH (n = 27)						
Time 1:	25 (93%)	21 (78%)	8 (30%)	4 (15%)	3 (11%)	1 (4%)
Time 2:	26 (96%)	25 (93%)	19 (70%)	10 (37%)	6 (22%)	2 (7%)

Table 2. Mean scores on background and summary ToM variables at Time 1 and Time 2, by group

Variable	Group	N	Mean	Std. Deviation
Time 1 Age	ASD	43	8.06	1.87

(years)	DoH	27	7.33	2.02
	TD	37	7.30	1.84
	Total	107	7.61	1.92
Time 2 Age (years)	ASD	43	9.48	2.07
	DoH	27	8.65	2.05
	TD	37	8.50	1.85
	Total	107	8.93	2.03
Language ability (z score)	ASD	37	-.04	1.08
	DoH	26	-.14	.80
	TD	24	.41	.85
	Total	87	.05	.96
Time 1 ToM Scale total (out of 6)	ASD	43	2.95	1.34
	DoH	27	2.30	.91
	TD	37	4.57	1.21
	Total	107	3.35	1.51
Time 2 ToM Scale total (out of 6)	ASD	43	3.67	1.32
	DoH	27	3.26	1.26
	TD	37	5.00	1.13
	Total	107	4.03	1.43
Total ToM (out of 8) at Time 1	ASD	43	3.58	1.98
	DoH	27	2.78	1.45
	TD	37	6.24	1.729
	Total	107	4.30	2.28
Total ToM (out of 8) at Time 2	ASD	43	4.65	1.90
	DoH	27	4.22	1.99
	TD	37	6.86	1.40
	Total	107	5.31	2.09
ToM Scale gain score	ASD	43	.72	.96
	DoH	27	.96	.81
	TD	37	.43	.73
	Total	107	.68	.86
Total ToM gain score	ASD	43	1.10	1.14
	DoH	27	1.44	1.37
	TD	37	.86	.97

	Total	107	1.12	1.18

Table 3. Numbers (and percent) of children (excluding those already at ceiling at Time 1) showing ToM progress, no change or decline

Group	Typically developing (TD) (n = 28)	Autism (ASD) (n = 42)	Deafness (DoH) (n = 27)
ToM Scale total (out of 6)			
Progress	15 (54%)	22 (52%)	19 (70%)
No Change	12 (43%)	18 (43%)	8 (30%)
Decline	1 (4%)	2 (5%)	0 (0%)
Total ToM (out of 8)			
Progress	17 (61%)	25 (60%)	20 (74%)
No Change	10 (36%)	16 (38%)	7 (26%)
Decline	1 (4%)	1 (2%)	0 (0%)

Table 4. Results (full model) of hierarchical multiple regression analysis exploring longitudinal predictors of Time 2 Total ToM score

Final Model (Block 3)	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	*t*	Significance (*p* value)
Age (Time 1)	.121	.087	.105	1.39	.168
Language Ability	.276	.155	.134	1.77	.080
Duration	.278	.305	.057	.91	.364
TotToM at Time 1	-.681	.070	.735	9.77	.000***
ASD	-.598	.394	-.150	-1.52	.133
Deafness	-.429	.427	-.100	-1.00	.318

Notes: Dependent variable = Time 2 TotToM score; Duration = months between Times 1 and 2;

*Significance:, *** denotes p < .001*

Figure 1

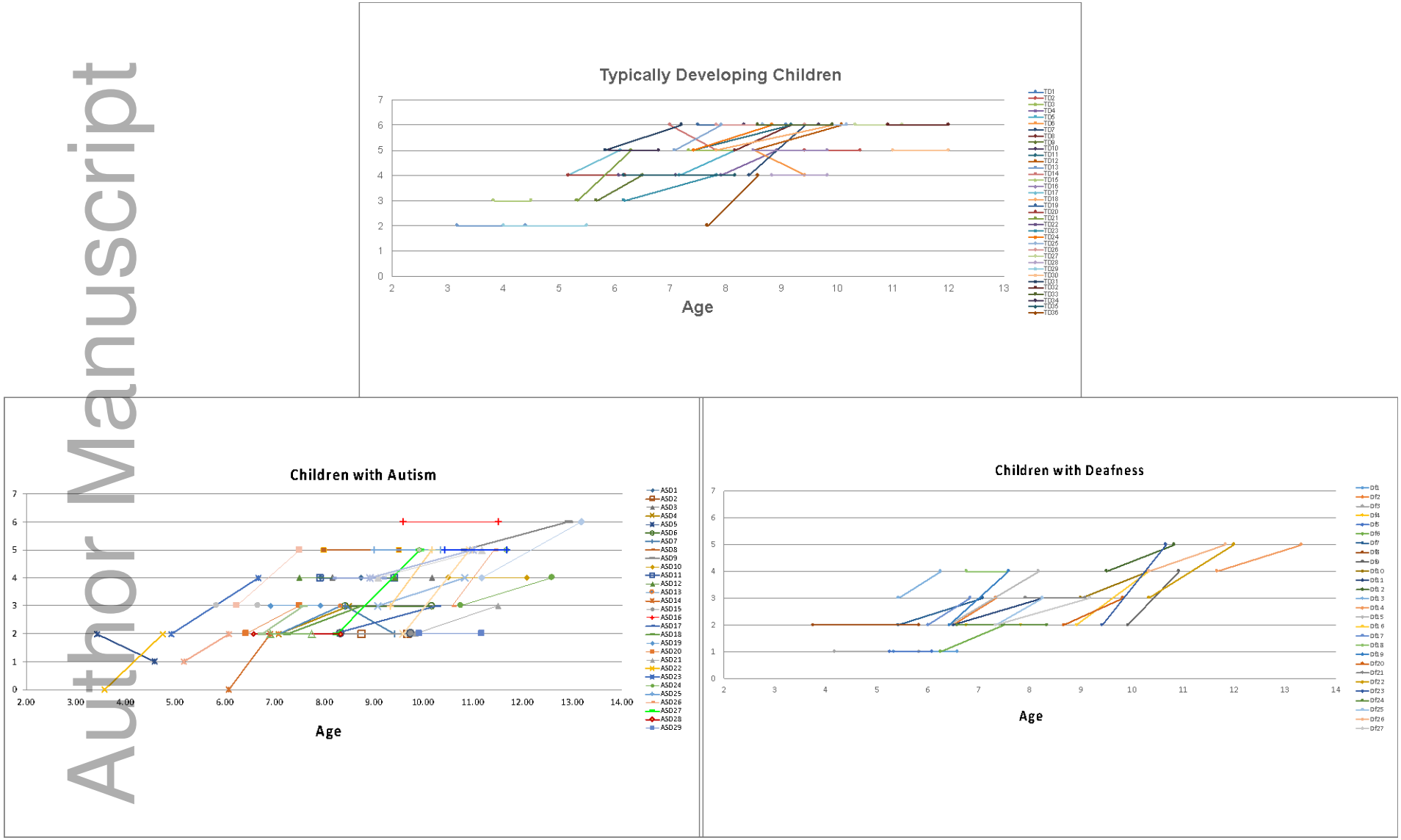

Figure 1. Longitudinal ToM Scale Progress for TD, ASD and DoH children