CHILD DEVELOPMENT

Child Development, May/June 2011, Volume 82, Number 3, Pages 780–792

Sequential Progressions in a Theory-of-Mind Scale: Longitudinal Perspectives

Henry M. Wellman
University of Michigan

Fuxi Fang
Chinese Academy of Sciences

Candida C. Peterson
University of Queensland

Consecutive retestings of 92 U.S. preschoolers (n = 30), Chinese preschoolers (n = 31), and deaf children (n = 31) examined whether the sequences of development apparent in cross-sectional results with a theory-of-mind scale also appeared in longitudinal assessment. Longitudinal data confirmed that theory-of-mind progressions apparent in cross-sectional scaling data also characterized longitudinal sequences of understanding for individual children. The match between cross-sectional and longitudinal sequences appeared for children who exhibit different progressions across cultures (United States vs. China) and for children with substantial delays (deaf children of hearing parents). Moreover, greater scale distances reflected larger longitudinal age differences.

Adults consistently interpret each other's actions in terms of underlying mental states (beliefs, desires, and emotions)—termed *theory of mind*—and children come to do so in the preschool years (Harris, 2006; Wellman, 2002). Theory of mind encompasses understanding of various mental states as well as how action is shaped by such mental states and experiences, not only in straightforward situations but also when mind and action are at odds because of forgetting, ignorance, false beliefs (FB), accident, and error. Thus, a ''standard'' way to assess theory-of-mind development is through FB tasks that require inferences about the action or thinking of someone whose beliefs conflict with reality and with the child's own current knowledge. Indeed, theory of mind is sometimes described as a preschool achievement equated with successful performance on FB tasks. However, we advocate a broader construal, both conceptually and developmentally. Achieving a theory of mind includes understanding multiple concepts acquired in developmental progression (Pons, Harris, & de Rosnay, 2003; Wellman & Liu, 2004). For this reason, researchers have recently established a Theory-of-Mind Scale (ToM Scale) and used it to examine the sequences of theory-of-mind understanding in different groups of children—typically developing children in several countries as well as deaf and autistic children who experience significant theory-of-mind delays (e.g., Peterson, Wellman, & Liu, 2005; Wellman, Fang, Liu, Zhu, & Liu, 2006; Wellman & Liu, 2004).

In brief, the ToM Scale provides a cross-sectional ordering of the developmental ease or difficulty of different theory-of-mind conceptions. A task battery such as this could sample from many mental state constructs and tasks, but this scale encompasses carefully constructed tasks assessing childhood understanding of (a) diverse desires (DD; people can have different desires for the same thing), (b) diverse beliefs (DB; people can have different beliefs about the same situation), (c) knowledge access (KA; something can be true, but someone might not know that), (d) FB (something can be true, but someone might

This research was supported by a grant from the U.S. National Institute for Child Health and Human Development (HD-22149) and by a grant from the National Natural Science Foundation of China (30270476), and the authors received indirect support from the University of Queensland and from the Australian Research Council. We gratefully acknowledge the helpful efforts of David Liu, Liu Yujuan, James Peterson, Kevin Brecker, Sara Parker, Kim Peterson, Jonathan Lane, and, especially, the children who participated with the generous consent of their parents.

Correspondence concerning this article should be addressed to Henry M. Wellman, Center for Human Growth & Development, University of Michigan, 300 North Ingalls 10th Floor, Ann Arbor, MI 48109. Electronic mail may be sent to hmw@umich.edu.

 0009-3920/2011/8203-0006
DOI: 10.1111/j.1467-8624.2011.01583.x

believe something different), and (e) hidden emotion (HE; someone can feel one way but display a different emotion). The tasks are devised to be similar in procedures, language, and format, yet U.S. preschoolers evidence a clear order of difficulty (as listed above), with understanding DD being easiest and understanding HE being hardest. This consistent progression has been confirmed by Guttman and Rasch scale analyses and in several corroborative studies with U.S. (e.g., Wellman, Lopez-Duran, LaBounty, & Hamilton, 2008) and Australian (e.g., Peterson et al., 2005) English-speaking preschoolers. Thus, the scale establishes a progression of conceptual achievements that pace theory-of-mind understanding in normally developing children, as well as a method for measuring that development.

Empirically, a scale can be formed from any collection of heterogeneous items as long as children only first pass some then successively pass some more. Theoretically, however, a scale progression is more valid and useful to the extent that it reflects an underlying conceptual progression or trajectory (Guttman, 1950). For this scale, the focal states, albeit different in many respects (e.g., feelings vs. knowledge), are arguably similar in being subjective and thus contrasting across individuals and with objective events or behaviors. That is, two persons can have contrasting desires for the same object or situation; similarly, they can have contrasting beliefs, or one can be knowledgeable where the other is ignorant. Relatedly, a person's mental state can contrast with behavior or with reality, as when a person feels one thing but expresses something different, or believes something not really true. Thus, conceptually, these contrasts all reflect the fact that mental states can be said to be subjective rather than objective in varying ways, and the scale addresses increasing steps in understanding mental subjectivity.

Patterns of success and failure, as revealed in developmental scales across children, do not definitively indicate that individual children proceed longitudinally through the identified sequences. Properly validated developmental scales reveal that the scaled tasks proceed from easiest to hardest and that cross-sectional groups of children systematically pass harder tasks at older ages. Thus they provide a cross-sectional approximation to developmental sequences. They potentially also can provide a cross-sectional shortcut to tracking longitudinal sequences. This potential, if validated, is of considerable note because complete longitudinal discovery and tracking of developmental sequences most often require large-scale, long-term studies that are costly, time consuming, and often noisy. The costs and time involved become particularly apparent if one considers delayed groups (e.g., children with autism or deafness) where, in the case of theory of mind, achievements that unfold for typically developing children from 2 to 6 years of age may require 12 years or more for such delayed children to accomplish.

Beyond the promise of a cross-sectional shortcut to longitudinal data for examining sequences of development, a scaling approach, and a validated scale, could provide a better measure of individual differences in theory of mind, by providing an extended metric of accomplishment instead of relying on a single attainment, such as FB alone. For now, measuring FB alone, albeit at times with a battery of varied FB tasks, is by far the most often used assessment of individual differences in preschool theory of mind (e.g., Hughes et al., 2005; Ruffman, Slade, & Crowe, 2002). Relatedly, a validated scale could also provide a more sensitive way to examine similarities and differences across groups or cultural communities, by examining their similar or different sequences of understandings, rather than just similarities and differences in average age of attainment for some one achievement.

Several studies now show the utility of scale scores and scale comparisons of this ToM Scale for comparing children across groups and cultures. The scale has been used informatively with children with autism and deafness (e.g., Peterson & Wellman, 2009; Peterson et al., 2005; Remmel & Peters, 2009), and has published, translated versions in German (Kristen, Thoermer, Hofer, Aschersleben, & Sodian, 2006) and Mandarin Chinese (Wellman et al., 2006), as well as unpublished versions in Japanese, Italian, Hebrew, and Korean. Several other studies indicate that it can provide a sensitive measure for researching individual differences (Aschersleben, Hofer, & Jovanovic, 2008; Wellman, Phillips, Dunphy-Lelii, & Lalonde, 2004). However, the crucial question remains of whether the sequences established cross-sectionally via the scale accurately portray the longitudinal developmental progressions that individual children undergo. In the current research, we address that question and provide evidence that (a) scale progressions and longitudinal progressions converge and that (b) groups who evidence different scale sequences evidence those sequential differences longitudinally as well. We establish these conclusions by examining the scale progressions of children given the scale at two or more successive times and we do so for three

groups of children—one from the United States, one from China, and a group of children born deaf to hearing parents.

These groups can provide a comprehensive longitudinal perspective on the sequences identified by the scale because they represent three different linguistic–cultural communities. U.S. and Chinese children are of comparative interest because of the contrasting language, cultural, and familial systems that characterize their early childhood experiences. Mainland Chinese children live in non-Anglo-European cultures, acquire non-Indo-European native languages, and live within cultural milieus often characterized as less individualistic and more collectivist or social-contextual than those of North America (Markus & Kitayama, 1991). Relatedly, they participate in distinctive cultural and linguistic parent–child practices (Tardif & Wellman, 2000) shaped by distinctive Confucian-Chinese meaning systems (Li, 2001; Nisbett, 2003) that differ from the Western European practices and meaning systems of middle-class, White Americans. On the hypothesis that social-interactive experiences and culturally shaped information critically influence theory-of-mind understandings, coupled with an analysis of focal language, cultural, and experiential differences, the theory-of-mind progressions for children growing up in China might well differ from those in Anglo-European communities. Indeed, these two groups cross-sectionally evidence two consistent, similar but crucially differing sequences of understanding on the ToM Scale (Wellman et al., 2006).

Children who are born deaf into hearing families provide an additional important and informative comparison. They too grow up within a different linguistic-cultural set of experiences. Despite valiant efforts to learn sign, hearing parents rarely achieve the proficiency of a native signer (Vaccari & Marschark, 1997). Hence, even if they themselves eventually master a signed language, deaf children in hearing families are unlikely to have anyone at home with whom to converse freely about mind-related topics like thoughts and beliefs. Moreover, these deaf children of hearing parents consistently show prolonged delays, and possibly incomplete development, for theory-of-mind understandings (Peterson, 2004; Peterson & Siegal, 2000; Schick, deVilliers, deVilliers, & Hoffmeister, 2007). Indeed, deaf children in Australia show the exact same five-step sequence on the ToM Scale but at substantially later ages than typically developing preschoolers in Australia and the United States (Peterson & Wellman, 2009; Peterson et al., 2005). Importantly, deaf children of hearing parents do not have the sorts of neurological impairments that characterize children with autism (who also show prolonged delays in theory-of-mind abilities). Furthermore, because deaf children of deaf parents (growing up with a signing parent) develop ToM on the hearing child's early timetable (Peterson & Siegal, 2000; Peterson et al., 2005), the theory-of-mind delays that arise when parents are hearing are more directly interpretable in terms of the impact of conversational–linguistic–interactive factors, factors that in more modest forms might impact all children as they come to understand persons' actions and minds.

In total, therefore, the current research can address several unanswered but critical questions. Focally, do cross-sectional theory-of-mind progressions established via the ToM Scale accurately depict sequences of understanding as they unfold in individual development? Moreover, are greater scale distances reflective of more age-related development? Are there some groups of children who never progress beyond some early or intermediate level of understanding? And, perhaps, are certain age periods representative of increased developmental theory-of-mind progress (e.g., perhaps for typically developing children centered around age 4 years; perhaps for children with delay due to deafness more centered around the transition to primary school)? To address these questions the current study encompasses a complementary mix of scaling along with longitudinal data for the same subjects. More generally, this novel mix of data can help provide needed information as to the sequences of conceptual understanding that characterize theory-of-mind development.

Method

Participants

Thirty-one Chinese preschoolers from Beijing, China who received the scale as 3-year-olds (M = 3-6, i.e., 3 years-6 months, range = 3-1 to 3-11) were retested as 4-year-olds (M = 4-6, range = 4-0 to 4-10) and 25 of them were retested a third time as 5-year-olds (M = 5-7, range = 5-1 to 6-0). Thirty U.S. preschoolers from a Midwestern university city who received the scale at one time (at ages ranging from 3-1 to 5-0, M = 3-11) were retested a second time, with retesting ranging from 6 months to 1½ years later (M = 13 months later). Thirty-one deaf children of hearing families from an Australian urban area, ranging in age from 4-2 to 12-8 (M = 8-3) at first testing, were tested a second time (M = 10-3) with delays from 8 months to almost

4 years (M = 24 months later), and 13 of these were tested a third or fourth time with final testing for 10 of them as teenagers, 12–16 years of age.

Although all the deaf children were severely or profoundly deaf late signers who used a sign language (rather than purely oral communication) in the classroom (and in our testing), 19 of the 31 had cochlear implants. This had no measurable impact on children's performance. Thus, a score representing total tasks passed (of the five) at their first testing session did not differ between deaf children with (M = 2.05) and without (M = 2.58) implants, $t(29) = 1.37$, $p > .15$.

Data from 12 of the 30 U.S. children for their first time of testing (Time 1) were previously included in Wellman and Liu (2004). Time 1 data for all of the 32 Chinese children were included in Wellman et al. (2006). Thirteen of the deaf children had their Time 1 data reported in Peterson et al. (2005) and a further 6 were included in Peterson and Wellman (2009). All data for later testings (Time 2, Time 3, and Time 4) are new to this report.

Tasks

Each child received the five-item scale (detailed in Wellman & Liu, 2004) at each testing. The five tasks are briefly described in Table 1: (a) DD, (b) DB, (c) KA, (d) contents FB, and (e) HE. These tasks all used toy figurines, and in one case (HE) a line drawing, for the target protagonists. Wellman, Cross, and Watson (2001) showed that for FB tasks, children answer very similarly when asked about real persons, dolls, toy figurines, story drawings, and real-life pictures or videos of persons.

Table 1
Brief Description of Tasks in the Scale

Task	Description
Diverse desires (DD)	Child judges that two persons (the child vs. someone else) have different desires about the same object
Diverse beliefs (DB)	Child judges that two persons (the child vs. someone else) have different beliefs about the same object, when the child does not know which belief is true or false
Knowledge access (KA)	Child sees what is in a box and judges (yes–no) the knowledge of another person who does not see what is in the box
Contents false belief (FB)	Child judges another person's false belief about what is in a distinctive container when child knows what is in the container
Hidden emotion (HE)	Child judges that a person can feel one thing but display a different emotion

Beyond using similar protagonists, the tasks were similar to one another in using picture props to show objects, situations, or facial expressions. These props helped present and remind children of the task contexts and response options. All the tasks were comparable in being based on, and asking about, a target contrast, for example, between one person's desire and another's, one person's perception and another's, a mental state (e.g., emotion or desire) versus a related behavior (e.g., an emotional expression or a choice of action). As a result, in each task there were two important questions asked: a target question about the protagonist's mental state or behavior and a contrast or control question about reality or expression or someone else's state. These consistent features gave all tasks a similar two-part presentation and a similar two-part format.

Essentially, children were tested with the exact same materials for each task at each time of testing. The primary exception was for deaf children and for the contents FB task. The deaf children were tested as often as four times and in some cases with delays between testings as short as 4 or 5 months. To eliminate the possibility these children might remember the specific deceptive container used for this task over such repeated testings, for deaf children different doll protagonists and different containers (e.g., a crayon box vs. a band aid box) were used at adjacent testings. This was not done for Chinese and U.S. children, who were tested less frequently and tested at consistently longer delays (94% of the time with delays of 11 or more months).

Task materials and wordings varied somewhat between the three groups, as appropriate for the different language systems (English, Mandarin, or Auslan and signed English) and national contexts (thus, the standard bandaid box used for contents FB in the U.S. and Australia was replaced with a familiar potato chip tube in China). Exact materials and wordings for the United States (Wellman & Liu, 2004), for the deaf children (Peterson et al., 2005), and for China (Wellman et al., 2006) are reported in earlier publications.

Procedures

Children were tested in a quiet room in their preschool by an adult experimenter. The tasks were presented in one of several orders. In all orders the DD task appeared early (as either the first or second task presented) to help children warm up to the process with a task hypothesized to be easier to understand. In all orders the HE task appeared last or next to last.

For the deaf children, two adults were present: an experienced male experimenter and one of several professionally trained interpreters of sign language who were highly familiar with the style of total communication used in each deaf child's classroom as well as with each child's own language preferences (e.g., for signed English vs. Auslan). Even at their initial test, the deaf children all had good everyday communication skills in this preferred language according to both their teachers' reports and their uniformly high levels of success on our control questions (which required similar lexical and syntactic competency to the test questions). Each interpreter was well known to the tested child and was employed in some capacity in the child's school. The interpreter, who was seated beside the experimenter and directly opposite and in full view of the participant, provided an accompanying translation of the experimenter's speech in the child's preferred mode of sign language, using a style of interpretation that was a familiar part of these children's everyday school routines. The interpreters paused while critical bits of stories were acted out (such as a doll's entry onto the scene), and both adults monitored that the child's gaze was directed at the props or the interpreter, as appropriate, before continuing each part of the procedure.

Scoring

All tasks included a focal test question as well as at least one other preliminary question or a control question or both. Like Wellman and Liu (2004), we ensured that children responded to the preliminary questions sensibly and attentively. In addition, we required that children pass any associated control questions, as well as test questions, in order to count as passing a task. Conversely, children were counted as failing a task if they failed the target test question or that question's associated control. This ensured that children comprehended and remembered all the relevant vocabulary, syntax and story information on which a meaningful, rather than random, response to a test question could be based. However, most children in all groups passed almost all of the control questions.

Results

Overview and Background

In Wellman et al. (2006), 80% of 135 U.S. and Australian English-speaking children's responses followed the sequence DD>DB>KA>FB>HE whereas 68% of 92 Chinese children followed the scale sequence DD>KA>DB>FB>HE. This difference—where Chinese children understood KA at an earlier step in their scale sequence in comparison to Anglo-Western children, who, in contrast, understood diverse beliefs earlier—was anticipated on the basis of focal language, cultural, and experiential differences, with the reasoning that, as argued by Nisbett (2003) and Li (2001), Western epistemology is focused more on truth and belief, whereas Chinese epistemology is focused more on pragmatic knowledge acquisition. Relatedly, there appear to be cultural differences such that Chinese children may receive more emphasis on ''knowing'' relative to ''thinking.'' As one example, in conversation with young children, Chinese parents comment predominantly on ''knowing'' (Tardif & Wellman 2000), whereas U.S. parents comment more on ''thinking'' (Bartsch & Wellman, 1995). Notably, however, in Wellman et al. (2006) both Chinese and English-speaking preschoolers scaled the same on a reduced four-item scale (DD>KA>FB>HE) that removed one item—86% of Chinese children and 88% of U.S. and Australian preschoolers scaled perfectly on that reduced four-item scale. Peterson et al. (2005) and Peterson and Wellman (2009) established that the sequence DD>DB>KA>FB>HE, which characterized U.S. and Australian English-speaking preschoolers, also characterized deaf Australian children of hearing parents but with delays on the order of 2 to 8 or more years at each of the five steps. In total, in past research approximately 65% to 90% of children in a group scale perfectly for five items, with more or less delay and with a critical difference for Chinese children in comparison to those from English-speaking Western societies.

Comparable patterns as to typical, alternative, and delayed sequences of development are evident in the current data—68% of the 31 Chinese children scaled perfectly (in the order DD>KA>DB>FB>HE) at Time 1, and 67% of the U.S. preschoolers scaled perfectly (in the order DD>DB>KA>FB>HE) at Time 1. Ninety percent of the deaf children scaled perfectly (DD>DB>KA>FB>HE) on their first session. Across all children and all sessions, 72% of the time that we gave this scale in the current study the U.S. children's responses fit the five-step pattern exactly, 96% of the time the deaf children's responses did so, and 79% of the time the Chinese children's responses fit their respective five-step pattern.

Guttman analyses confirm that these children's patterns were generally scalable. For this overall confirmation, we analyzed data for all 92 children at

their first and again at their second times of testing (because all children were tested twice but many fewer had three, and still fewer had more than three testings). For this analysis, U.S. hearing and Australian deaf children were rated in terms of the sequence DD>DB>KA>FB>HE, and Chinese children were rated in terms of DD>KA>DB>FB>HE as appropriate given both initial descriptive data and past cross-sectional research. Green's (1956) index of reproducibility for this combined data (across all groups and two times of testing) was *Rep* = .95 (values > .90 indicate scalable items), and his more conservative index of consistency (which compares observed patterns to responding expected by chance) was *I* = .52 (values above .50 are significant). Thus, just as in prior research most children fit their scale sequences at first and second (and indeed all) times of testing.

Moreover, just as in past research, the U.S., Australian, and Chinese children all fit an identical reduced four-item scale sequence (DD>KA>FB>HE). As can been seen in Table 2, across their first two testings, these U.S. children did so 88% of the time, these Chinese children did so 85% of the time, and these deaf children did so 94% of the time. (For all three groups this common four-step scale pattern was significant both in terms of reproducibility and consistency; *Rep*s = .97, .96, .99, *I*s = .61, .50, .87.)

Longitudinal Progressions

Given scalable results overall, a succinct way to summarize the longitudinal data is in terms of children's overall scale scores (0–5 possible items correct). Figure 1 graphically presents the longitudinal progressions for these five-item scale scores. As that figure conveys, in all groups children generally increased longitudinally and decreases (going backward) were rare. In fact, from Time 1 to Time 2, no U.S. child's score decreased, 24 of 30 (80%) increased, and 6 stayed the same. From Time 1 to Time 2, 22 of 31 (71%) Chinese children's scores increased, 5 stayed the same, and 4 decreased. From Time 1 to Time 2, 1 deaf child's score decreased, 19 of 31 (61%) increased, and 11 stayed the same. In total, on only 7 of 133 total consecutive testings (or 5% of the time) did children ''regress,'' and a high majority of children increased.

Table 3 summarizes the mean scale score data. Inferential analyses confirm the graphical and summary data. A 2 (first vs. second testing) × 3 (groups: U.S., Chinese, deaf) repeated measures analysis of variance (ANOVA) yielded main effects of group, $F(2, 89) = 6.14$, $p < .01$, $\eta^2 = .12$, and of first–second testing, $F(1, 89) = 110.30$, $p < .001$, $\eta^2 = .55$. A parallel 2 × 3 ANOVA on each child's first and last testing similarly showed main effects of group, $F(2, 89) = 5.48$, $p = .01$, $\eta^2 = .11$, and of first–last testing, $F(1, 89) = 209.52$, $p < .001$, $\eta^2 = .70$. Considering first versus final testing includes the largest extent of delay between the testings, and all children had a first and last test (albeit the last testing might be the second, third, or even fourth for differing children). These analyses thus confirm the patterns in Figure 1; children in all groups mostly proceeded up the scale as they grew older. Post hoc (Tukey honestly significant difference) tests confirm that the first–second data are not significantly different for the U.S. and Chinese children, but that both these groups significantly differ from the deaf children (*p*s < .03). For all groups, however, scores

Table 2
Guttman Scalogram Patterns for a Reduced Four-Item Scale

Pattern	1	2	3	4	5		
Diverse desire (DD)	−	+	+	+	+		
Knowledge access (KA)	−	−	+	+	+		
Contents false belief (FB)	−	−	−	+	+		
Hidden emotion (HE)	−	−	−	−	+		
						Other patterns	Total
U.S. (*n* = 30)	0	12	9	20	12	7	60
China (*n* = 31)	0	9	20	16	8	9	62
Deaf (*n* = 31)	4	31	10	9	5	3	62
Total	4	52	36	46	24	22	184

Note. A minus means a child failed the task in question; a plus means the child passed. The five focal patterns represent 5 of the total possible 24 patterns of response encompassing the four dichotomous items. A child evidencing any of the remaining 19 patterns was classified as Other. The complete data would include five items and six scaled patterns; however, the item orders would be DD>DB>KA>FB>HE for the U.S. and deaf children but DD>KA>DB>FB>HE for the Chinese children.

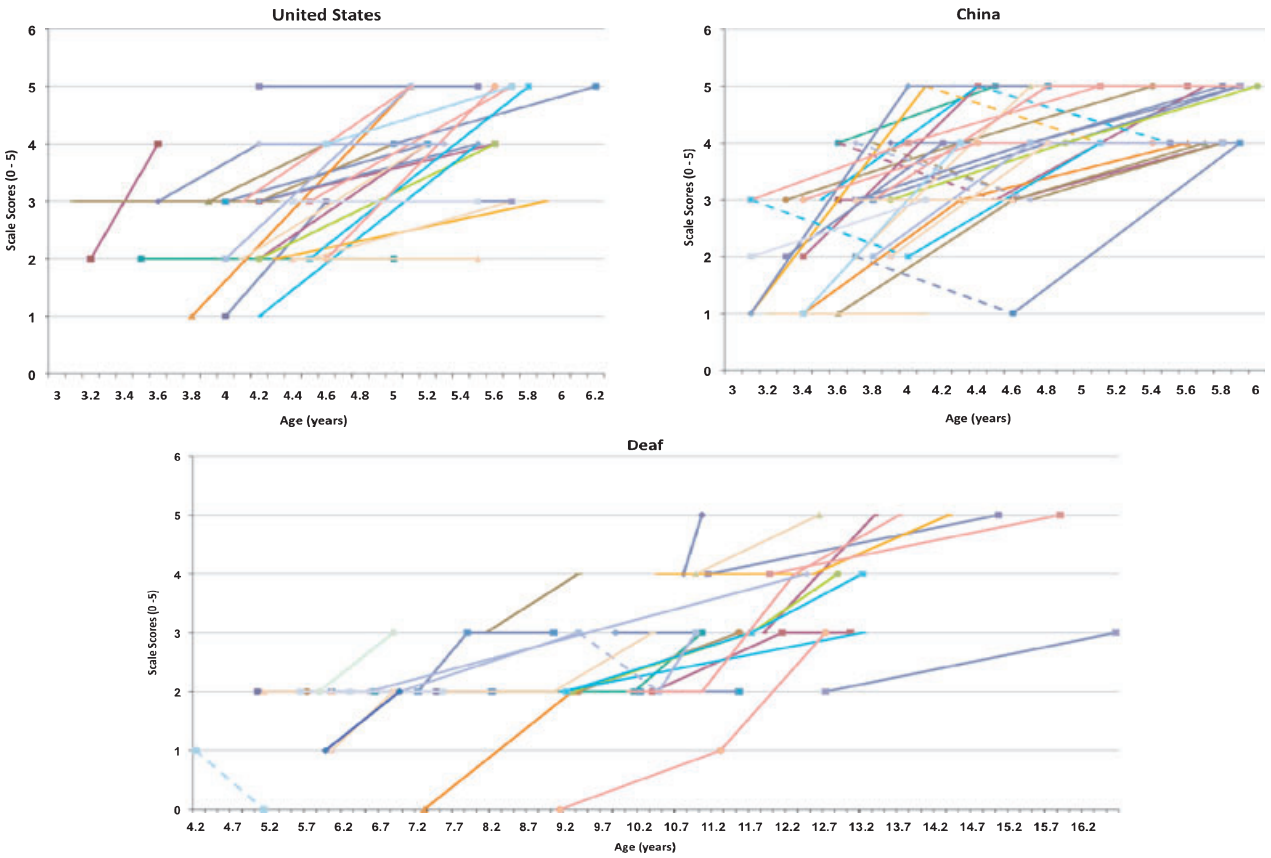

Figure 1. Panels showing longitudinal changes in total scale scores (0–5) for individual children in the three groups: U.S. children at the top left, Chinese children at the top right, and Australian deaf children of hearing parents across the bottom.

Table 3
Average Scores and Means Ages at the Different Times of Testing

Mean age	Time 1	Time 2	Time 3	Time 4
U.S. average score				
4.14	2.59 (n = 30)			
5.26		3.97 (n = 30)		
China average score				
3.54	2.52 (n = 31)			
4.49		3.68 (n = 31)		
5.64			4.52 (n = 25)	
Deaf average score				
8.22	2.19 (n = 31)			
10.23		2.87 (n = 31)		
11.00			3.08 (n = 13)	
11.30				3.33 (n = 3)

increased from first to second testings (ps < .03) and from first to last testings (ps < .03).

As can be seen in Table 3, the average numbers of steps advanced from Time 1 to Time 2 were 1.38 for U.S. children, 1.16 for Chinese children, and 0.68 for deaf children. Thus, deaf children evidence the least amount of change despite having delays averaging 24 months between their first and second testings. Cross-sectional data from earlier studies predict that progress would be slower for deaf children; to capture a more complete picture in their case we were able to test 13 of these children three or more times with delays from first to last testing ranging from 2 to 6 years. For these 13 deaf children, increases averaged 1.31 steps over a time span averaging 41 months, similar to a 1-year change for the U.S. and Chinese hearing children.

Cross-sectional scale data alone provide a metric of change by calculating average ages for children with scores of 0, 1, 2, 3, 4, or 5. As shown in the top half of Table 4, average age increased across these accumulating scores in each group, but additionally average ages were very different across the groups. Specifically, deaf children were delayed in age at each step in comparison to both typically developing groups. Thus, a 6 (score: the six different possible scores) × 3 (group: U.S. vs. Chinese vs. deaf

Table 4
Average Ages of Children for the Increasing Scores on the ToM Scale

Scores from current data	0	1	2	3	4	5
U.S.: Average ages (n = 31)	—	4.07	4.18	4.35	5.14	5.53
SD		0.23	0.43	0.59	0.57	0.38
China: Average ages (n = 31)	—	3.56	3.61	3.93	4.88	5.26
SD		0.53	0.31	0.52	0.74	0.67
Deaf: Average ages (n = 31)[b]	7.12	6.85	7.95	10.98	11.60	13.61
SD	2.0	3.1	2.0	2.3	1.2	1.4
Comprehensive data across several studies[a]						
U.S. and Australian preschoolers (N = 280)	3.22	3.66	3.84	4.45	4.77	5.15
Chinese preschoolers (N = 135)	—	3.39	3.74	4.14	4.97	5.19
Deaf children of hearing[b] parents (N = 66)	8.77	7.83	7.92	9.88	11.31	12.40

[a]Data for U.S. and Australian preschoolers were obtained from Wellman and Liu (2004), Wellman, Phillips, Dunphy-Lelii, & Lalonde (2004), Peterson & Wellman (2009), and Peterson and Wellman (2009). Data for Chinese preschoolers came from Wellman, Lopez-Duran, LaBounty, & Hamilton, (2008). Data for deaf children of hearing parents are from Peterson, Wellman, and Liu (2005) and Peterson and Wellman (2009). [b]As is clear in Table 4, deaf children of hearing parents often show a pattern where average age of those children who fail all items (score = 0) is higher (rather than lower) than children who pass one or two items. Presumably, deaf children receiving 0s include some younger children who are only beginning their ToM understandings, and also older children who evidence more protracted difficulties in acquiring ToM understandings. The current longitudinal data, however, demonstrate that individual children, including deaf individuals, almost always make progress in their ToM understandings and almost never go backward.

groups) ANOVA with average age as the dependent measure, yielded significant main effects of score, $F(5, 209) = 29.38$, $p < .001$, $\eta^2 = 0.41$, and group, $F(2, 209) = 330.83$, $p < .001$, $\eta^2 = .76$, subsumed under a significant interaction, $F(8, 209) = 9.55$, $p < .001$, $\eta^2 = .27$. As shown in Table 4, variances are also considerably greater for the deaf children than for the U.S. or Chinese hearing ones. Thus, progressions on this scale are substantially slower for deaf children who are also more variable in the speed and extent to which they make progress. Nonetheless, univariate ANOVAs within the U.S., Chinese, and deaf groups confirm that all groups progress: For U.S. children as scores increase from 1 to 2 to 3 to 4 to 5, average ages consistently increase, $F(4, 55) = 16.74$, $p < .001$, $\eta^2 = .55$, as holds true for Chinese, $F(4, 82) = 23.20$, $p < 001$, $\eta^2 = .53$, and deaf, $F(5, 72) = 17.23$, $p < .001$, $\eta^2 = .54$, children as well. More importantly, the current longitudinal analyses confirm that as individual children get older their scores consistently increase. As we show next, children's scores not only increase, they increase in sequence along the scale.

Perfect Scaling Progressions

Almost all children's responses closely approximate the overall scale, and the vast majority fit their scale pattern exactly at each time of testing (75% of children fit the scale exactly at Time 1). For those children it is especially easy and revealing to examine their longitudinal progressions in more detail by identifying how many proceed along the scale in order, as opposed to going backward, or skipping a step by passing a harder step out of order first. For this analysis, children who scored identically at two times were counted as proceeding in order (they evidence perfect scale patterns and neither go backward nor skip). Eighteen of 20 U.S. children proceeded in order from Time 1 to Time 2 (with none going backward and 2 skipping over an incorrect answer to pass a harder answer out of order first), 16 of 21 Chinese children proceeded in order from Time 1 to Time 2 (with 3 going backward and two skipping), 27 of 28 deaf children proceeded in order (with 1 skipping by passing a harder step out of order first). In short, the scale progressions capture children's individual progress very well; 88% of these children either maintained the same score (21%) or proceeded longitudinally in order (67%). Similar numbers accrue if calculating the progressions from Time 1 to each child's last time of measurement. In sum, most individual children proceeded in order up the scale; progressions that skip ahead (8% of the total possible) are infrequent and those going backward (4%) constitute a number so small that it may simply represent measurement error.

Extent of Progression

Because delay of testing from one time to the next varied, we could examine the consistent

longitudinal progressions (shown in Table 2 and Figure 1) in still further detail by considering each child's changed score from their first to last testing as related to their age difference between first and last testing. (Ages were more delayed and differences more varied from first to last testing than for first to second testing, yielding more variance for the correlations. For U.S. children first and last testing is the same as first and second testing because each child had only the two testings; consequently, variances were more compressed for the U.S. children.) For the U.S. children, change in age (from Time 1 to Time 2) was uncorrelated with change in score (increase in scale score from Time 1 to Time 2), $r(n = 30) = .10$. But for Chinese children, change in age from first to last testing correlated significantly with change in score, $r(n = 31) = .36$, $p < .05$. And for deaf children, these correlations were significant as well, $r(n = 31) = .43$, $p < .02$. Thus, not only do individual children's scores at Time 2 and Time 3 increase over Time 1, moreover, when delays are more extensive (as in the delays of 2–5 years encompassed by first to last testing for Chinese and deaf children), then variation in amount of delay predicts amount of score increase; more time between testing results in larger score increases.

Conceivably, progress on these ToM items (which were specifically chosen to represent preschool acquisitions for typically developing children) could stop in the preschool years if unachieved then. However, as shown in Figure 1, deaf children were continuing to make progress in the age range from 10 to 16 years. Indeed, for these deaf children of hearing parents, middle childhood seems an age of special theory-of-mind progress. Figure 1 shows sizable increases in performance between ages 10 and 13, with no child in this longitudinal sample achieving the final step on the scale (understanding HE) until after age 11. The cross-sectional data in Table 4 help confirm that the final steps on the ToM scale (represented by scores of 4 and 5) are not achieved by many deaf children of hearing parents until the age of 11 or 12 years. In sum, data from late-signing deaf children demonstrate that progress on these ToM understandings if unachieved early in life can nonetheless be extensive in later childhood.

Longitudinal Predictions

Optimally, in a cognitive developmental sequence, earlier understandings should not just precede later understandings but also longitudinally predict them. We examined this issue by considering whether children's earlier performance on the easier tasks predicted their later performance on the harder tasks. In the present data the largest number of comparable cases are children for whom we have data both essentially as young 4-year-olds (range = 4-0 to 4-9) and then again later as 5-year-olds (range = 5-0 to 6-0) with testing approximately 1 year apart. Forty-six U.S. ($n = 21$) and Chinese ($n = 25$) children provide data in this range. (It would be uninformative to simply compare first and second testings across all children because age at either testing would vary so widely—e.g., from 3-1 to 12-8 at first testing—that it would completely confound and render uninterpretable an ''early–later'' comparison.)

In this analysis, easier, earlier items were the sum of each child's performance on DD + DB + KA (at 4 years) and harder, later items were the sum of each child's performance on FB + HE (at 5 years). Although U.S. and Chinese children have a reversal of sequence between DB and KA, for both groups DD, DB, and KA are easier and earlier than FB and HE. We used composites of 2 (FB + HE) or 3 (DB, DD, KA) items to have scores with more extended variance for these correlational analyses. Children's performances on the easier items at the earlier age indeed predicted their performance on the harder items approximately 1 year later, $r(n = 46) = .41$, $p < .005$.

Discussion

In general, children longitudinally progress through the five tasks in this Theory-of-Mind Scale in a standard order. Moreover, their longitudinal progression matches the same order obtained in cross-sectional data that ranks the tasks from easiest to hardest (i.e., from tasks passed by the largest number of children to the fewest number). These data thus confirm the cross-sectional scores and scalings as valid approximations to underlying longitudinal developmental progressions; this particular scale represents a useful cross-sectional shortcut for examining sequences of theory-of-mind understanding. This conclusion and these findings are strengthened because they hold across three disparate and contrasting cultural groups that evidence somewhat different sequential progressions (e.g., United States vs. China) and that encompass typical preschool progressions but also extensive delay (U.S. and Chinese preschoolers vs. deaf children of hearing parents).

Our results represent more than just a needed methodological validation; they confirm and extend several substantive findings as well. As a cornerstone of social intelligence and satisfying social interaction, theory of mind develops rapidly in the preschool years, and our focus concerns theory-of-mind insights that arise for normally developing children within the preschool years. Conceivably, all these mental-state insights might be equally hard for children: All concern subjective, internal states (desires, ignorance, beliefs, feelings) that are potentially at odds with overt behavior or external reality. Equally conceivable, children might understand some states before others, but earlier-understood versus later-understood states would not be consistent from one child to the next, depending on different individual experiences or family foci of conversations (e.g., emotions vs. wants vs. ignorance). In contrast to either of these alternatives, our data confirm distinct regularities in children's developing understanding of mind.

One issue for scaling findings is the possibility that the task order obtained reflects more a logical dependence between tasks, rather than a psychological-developmental progression. Perhaps, Task B (measuring *b*) succeeds Task A (measuring *a*) only because *b* is logically composed of *a* plus something more (Brandtstadter, 1987). However, our results speak against mere logical dependence because for U.S. children DB is easier and comes before KA (see also Wellman & Liu, 2004), but for Chinese children KA is easier and comes before DB (see also Wellman et al., 2006). The fact that order of difficulty reverses for these tasks across groups argues against the possibility that one task or the other was intrinsically more difficult logically, or simply more difficult in terms of task demands or linguistic complexity. Similarly, other research has shown that FB is understood before HE for most children but that for individuals with autism the order reverses to become HE before FB (Peterson et al., 2005).

These reversed and alternative sequences also make it unlikely that the focal theory-of-mind progressions simply represent childhood increments in executive function or cognitive complexity, manifest, perhaps, because these scale tasks increase such demands step by step (Andrews & Halford, 2002; Frye, Zelazo, & Palfai, 1995). To the contrary, these ToM Scale tasks arguably place very similar demands on executive function and require similar levels of cognitive complexity: All deal with two alternatives, and one alternative must be inhibited to correctly choose the other (e.g., diverse desires task: inhibit my preference, answer on the basis of another person's preference; diverse beliefs task: inhibit my belief, answer on the basis of another person's belief; KA task: inhibit my knowledge or reality, answer on the basis of another person's knowledge). More crucially still, suppose one were to devise more subtle task analyses to claim that there are step-by-step increments in executive functioning or complexity demands across these tasks (as has been done for a comparison between perspective taking, appearance reality, and FB by Andrews, Halford, Bunch, Bowden, & Jones, 2003). Any such analyses proposed to account for the detailed U.S. data (e.g., that KA requires more inhibition or more complex reasoning than DB) would be challenged by the Chinese data. Any such analyses proposed to account for the progression (amid delay) for deaf children of hearing parents would be challenged by the alternative progression (amid delay) evident for children with autism.

Nothing in our data or this discussion should be taken to conclude that domain-general preschool advances in executive functioning or in cognitive complexity are unimportant to children's developing theories of mind. Both are known, potent influences (Halford, Cowan, & Andrews, 2007). The current discussion and data only argue that these domain-general cognitive changes are not the full story; conceptual insights specific to subjective, mental state understandings are also required for the progressive attainment of theory of mind.

Because our focus on theory of mind is centrally concerned with advances as measured by preschool judgment tasks, the scale methods and tasks are akin to ''standard'' FB tasks, albeit broader in conceptual content. Theory-of-mind development begins in infancy, and some recent research claims infants—at 12 to 15 months—already have an awareness that actors act on the basis of their beliefs and FB (e.g., Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007). It is not yet clear how to best interpret these infant ''FB'' findings or how to reconcile or integrate them with the preschool ones. Regardless, something definite and important is happening in children's theory-of-mind understandings in the preschool years, beyond earlier developments in infancy. Differences in FB understanding as measured in the preschool years, for example, predict several key childhood competences, such as how and how much children talk about people in everyday conversation, their engagement in pretense, their social interactional skills, and consequently their interactions with and popularity with peers (Astington

& Jenkins, 1995; Lalonde & Chandler, 1995; Watson, Gelman, & Wellman, 1998). Variability in preschool performance on theory-of-mind tasks overlaps with but is distinctively different from executive function and IQ advances during the preschool years (e.g., Carlson & Moses, 2001; Wellman et al., 2008). These empirical connections and findings are important for confirming theory of mind's significance and relevance during the preschool years as indexed by preschool theory-of-mind tasks (especially as researched thus far for FB tasks). Our data now show that these important preschool theory-of-mind understandings proceed in a progressive conceptual sequence from those mastered early to those achieved substantially later.

It is important to emphasize that the empirical fact of statistically reliable sequentiality does not necessarily imply a cause–effect relation between earlier and later steps in the ToM sequence. Our data do not confirm, for example, that earlier desire understandings are necessary in shaping later theory-of-mind understanding (such as FB or HE). Even longitudinal evidence, including predictive longitudinal relations of the sort we report cannot fully provide that confirmation; systematic training or experimental studies are needed to establish such causal conclusions. However, careful data as to consistent sequences (among carefully chosen, comparable tasks within a targeted domain of understanding) helpfully inform and constrain such experimental research. If understanding A reliably succeeds understanding B, it is implausible that it shapes and causes B; if it reliably precedes understanding B, it becomes a plausible causal candidate for further research. In this vein, the present five-step ToM scale could prove to be a particularly promising empirical tool for designing training research, aimed at theoretical or practical concerns. The scale provides a blueprint of sequenced understandings that could be used to build on one another and provides a reliable, extended measuring device for assessing systematic gains. The spread of the scale could prove especially advantageous for assessing short-term improvements in delayed groups (like late-signing deaf children or those with autism) who may require over 10 years to achieve the traditional ToM criterion (FB) if untutored development is allowed to run its course.

More generally, use of this Theory-of-Mind Scale (and validated scales more generally) shortcuts discovery and analysis of these progressions, progressions that would take many years to emerge in longitudinal research. Cross-sectional scale results are not only revealing in their own right, they aid in the effective, well-targeted design of longitudinal research as well. Moreover, they allow the conduct of research that informatively combines cross-sectional and longitudinal perspectives on developmental trajectories, such as the current study.

Including deafness in investigations of theory of mind seems to us especially informative. For deaf children of hearing parents, like typically developing children, an understanding of desire precedes corresponding understanding in the realm of belief. In addition, an initial understanding of knowledge and ignorance develops ahead of the understanding that someone can hold a belief that the child knows is decidedly false. Yet, at each of these steps late-signing deaf children come to these understandings at ages that are older, by several years, than their age of acquisition by typically developing children (and by native signers; see Peterson et al., 2005). A protracted sequence like this on a trajectory of steps common with that of typically developing children can help researchers examine progressions that typically crowd together very rapidly but nonetheless sequentially influence each other. Moreover, because deaf children do not have the central neurological deficits that characterize children with autism, but rather peripheral auditory deficits, their progressions more clearly and precisely reflect the influences of, and departures from, the normal course of experience with social interaction, language, and conversation. Given (a) the known impact of such language-saturated factors for developing theory of mind in the normal case (Astington & Baird, 2005) and (b) the specific deficits in just these factors that characterize deaf children of hearing parents, research specifically with deaf children can critically address how social–interactive–language factors influence a cascade of understandings about minds in connected ways, as demonstrable in a delayed, yet consistent, progression of theory-of-mind understanding.

Because of their atypical conversational experiences and delays, for example, deaf children of hearing parents provide important means for researching questions about critical periods of development (e.g., Newport, 1991). A critical period hypothesis for theory-of-mind development (Siegal & Varley, 2002) could conceivably predict that those children (e.g., deaf children of hearing parents) who missed out on specific early conversational inputs (e.g., discussions with family members' of mental states such as knowledge and beliefs) during a crucial preschool period might be forever blocked in developing beyond some early

level of understanding. Thus, they might be unable to proceed beyond some early or intermediate point on our ToM scale. The current longitudinal data suggest to the contrary that children can continue to make substantial progress on these ''preschool'' theory-of-mind insights well into adolescence. Virtually all of the late-signing deaf children in this sample were continuing to progress longitudinally through the scale at advanced ages. Pyers and Senghas (2009) recently reported related data. Deaf Nicaraguan adults continued to improve on FB understanding (and increasingly used mental-state discourse) over a 2-year period as adults in their 20s.

Several limitations and clarifications of the current research deserve mention. The specific scale and scale items tested here represent a single, inevitably limited assessment of theory-of-mind progressions. The items were carefully selected and devised to have the advantages of being comparable in testing format, few in number, and easily understood by young children (to facilitate use with preschoolers and those with developmental delays). Moreover, the five items that form this ToM Scale were specifically chosen because they form a strict Guttman Scale where if a child passes a ''later'' item he or she consistently passes all ''earlier'' ones as well. As a consequence of these choices, these tasks do *not* capture additional important aspects of theory-of-mind development, do not capture all important theory-of-mind progressions, do not encompass all preschool theory-of-mind insights, and moreover ''ceiling out'' for use with older children who continue to achieve further theory-of-mind progress. Additional items and approaches are informative; as an example, Pons et al. (2003) provide a battery of items assessing childhood understanding of emotional states (largely neglected in the current scale) and using a complementary, non-Guttman approach to scale construction.

The current longitudinal data are specifically limited in that for some children they rest on only two testing occasions and for most others only three. Increased longitudinal assessments could better capture individual children's progression across each of these five tasks, approximating the densely sampled longitudinal trajectories sometimes possible in microgenetic research. Nonetheless, our mix of differing numbers of testing occasions coupled with differing mixes of testing delays encompassing groups with quite different rates of development provides a valuable and comprehensive initial longitudinal picture.

In total, the theory-of-mind scale validated here establishes (a) a progression of conceptual achievements that mark social cognitive understanding in normally developing preschool children, (b) a similar progression that marks delayed though continuing conceptual achievements for deaf children of hearing parents, and (c) a method for measuring that development accurately and informatively that provides (d) a good approximation to longitudinal data via a cross-sectional approach.

References

Andrews, G., & Halford, G. S. (2002). A complexity metric applied to cognitive development. *Cognitive Psychology*, *45*, 153–219.

Andrews, G., Halford, G. S., Bunch, K. M., Bowden, D., & Jones, T. (2003). Theory of mind and relational complexity. *Child Development*, *74*, 1476–1499.

Aschersleben, G., Hofer, T., & Jovanovic, B. (2008). The link between infant attention to goal-directed action and later theory of mind abilities. *Developmental Science*, *11*, 862–868.

Astington, J. W., & Baird, J. A. (2005). *Why language matters for theory of mind*. New York: Oxford University Press.

Astington, J. W., & Jenkins, J. M. (1995). Theory of mind development and social understanding. *Cognition and Emotion*, *9*, 151–165.

Bartsch, K., & Wellman, H. (1995). *Children talk about the mind*. New York: Oxford University Press.

Brandtstadter, J. (1987). On certainty and universality in human development: Developmental psychology between apriorism and empiricism. In M. Chapman & R. A. Dixon (Eds.), *Meaning and the growth of understanding* (pp. 69–84). New York: Springer-Verlag.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*, 1032–1053.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, *10*, 483–527.

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

Guttman, L. (1944). A basis of scaling quantitative data. *American Sociological Review*, *9*, 139–150.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Science*, *11*, 236–242.

Harris, P. L. (2006). Social cognition. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology* (5th ed., pp. 811–858). New York: Wiley.

Hughes, C., Jaffe, S. R., Happe, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of indivdual differences

in Theory of Mind: From nature to nurture? *Child Development*, *76*, 356–370.

Kristen, S., Thoermer, C., Hofer, T., Aschersleben, G., & Sodian, B. (2006). Skalierung von ''theory of mind'' aufgaben [Scaling of theory of mind tasks]. *Zeitschrift fur Entwicklungspsychologic und Padagogische Psychologie*, *38*, 186–195.

Lalonde, C. E., & Chandler, M. J. (1995). False belief understanding goes to school: On the social-emotional consequences of coming early or late to a first theory of mind. *Cognition and Emotion*, *9*, 167–185.

Li, J. (2001). Chinese conceptualization of learning. *Ethos*, *29*, 111–137.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224–253.

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

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently—and why*. New York: Free Press.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Peterson, C. C. (2004). Theory-of-mind development in oral deaf children with cochlear implants or conventional hearing aids. *Journal of Child Psychology & Psychiatry*, *45*, 1–11.

Peterson, C. C., & Siegal, M. (2000). Insights into theory of mind from deafness and autism. *Mind & Language*, *15*, 123–145.

Peterson, C., & Wellman, H. M. (2009). From fancy to reason: Scaling deaf and hearing children's understanding of 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 and deafness. *Child Development*, *76*, 502–517.

Pons, F., Harris, P. L., & de Rosnay, M. (2003). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European Journal of Developmental Psychology*, *2*, 127–152.

Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science*, *20*, 805–812.

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

Siegal, M., & Varley, R. (2002). Neural systems involved in theory of mind. *Nature Reviews Neuroscience*, *3*, 462–471.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*, 580–586.

Tardiff, T., & Wellman, H. M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology*, *36*, 25–43.

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.

Watson, J. K., Gelman, S. A., & Wellman, H. M. (1998). Young children's understanding of the non-physical nature of thoughts and the physical nature of the brain. *British Journal of Developmental Psychology*, *16*, 321–335.

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, UK: 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 understanding in Chinese children. *Psychological Sciences*, *17*, 1075–1081.

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

Wellman, H. M., Lopez-Duran, S., LaBounty, J., & Hamilton, B. (2008). Infant attention to intentional action predicts preschool theory of mind. *Developmental Psychology*, *44*, 618–623.

Wellman, H. M., Phillips, A. T., Dunphy-Lelii, S., & LaLonde, N. (2004). Infant social attention predicts preschool social cognition. *Developmental Science*, *7*, 283–288.