Child Development, March/April 2004, Volume 75, Number 2, Pages 523–541

Scaling of Theory-of-Mind Tasks

Henry M. Wellman and David Liu

Two studies address the sequence of understandings evident in preschoolers' developing theory of mind. The first, preliminary study provides a meta-analysis of research comparing different types of mental state understandings (e.g., desires vs. beliefs, ignorance vs. false belief). The second, primary study tests a theory-of-mind scale for preschoolers. In this study 75 children (aged 2 years, 11 months to 6 years, 6 months) were tested on 7 tasks tapping different aspects of understanding persons' mental states. Responses formed a consistent developmental progression, where for most children if they passed a later item they passed all earlier items as well, as confirmed by Guttman and Rasch measurement model analyses.

Children's understanding of persons' mental states —their theory of mind—is a crucial cognitive development and has been intensely studied in the last 15 years (e.g., see Flavell & Miller, 1998). At times, theory of mind is discussed as a single cognitive process or achievement (especially in some areas of inquiry, such as primate cognition or research on autism). Relatedly, much theory-of-mind research has focused on a single task paradigm examining children's understanding of false belief. However, many researchers believe that developing a theory of mind includes understanding multiple concepts acquired in an extended series of developmental accomplishments (for a recent review, see Wellman, 2002). Consequently, investigations of young children's understandings of intentions, emotions, desires, knowledge, and other states have become prevalent. However, little research empirically establishes developmental progressions in children's various understandings. Support for one progression comes from studies showing that children's understanding of desires seems to precede their understanding of beliefs (e.g., Bartsch & Wellman, 1995; Flavell, Flavell, Green, & Moses, 1990; Gopnik & Slaughter, 1991; Wellman & Woolley, 1990). But other progressions are empirically unclear or contentious (e.g., Mitchell, 1996; Perner, 1995). More serious still, very little research has attempted to investigate comprehensively an extended series of theory-of-mind developments.

We assume that, for normally developing children, certain insights about the mind develop in a predictable sequence. We hypothesize that these insights index an underlying developmental progression that could be captured in a theory-of-mind scale. We provide two types of empirical support for this hypothesis. First, we report a preliminary meta-analysis of studies that have compared different types of mental state understandings (e.g., desires vs. beliefs or ignorance vs. false belief). A meta-analysis seems useful to integrate and clarify scattered individual findings that are at times contradictory. Primarily, however, we report a study testing a theory-of-mind scale for preschool children—a set of methodologically comparable tasks that focus on differing conceptual constructs that may developmentally appear in sequence.

As background, our focus concerns preschool developments, a developmental period when there are many changes in mental sate understanding. We do not include second-order false-belief tasks (which regularly are acquired in the early school years, consistently after a first-order understanding of false belief; Perner & Wimmer, 1985), nor do we include tasks representing more mature (Wellman & Hickling, 1994) or advanced theory-of-mind understandings (Happe, 1994) thought to be acquired later in development and that focus largely on metaphor, irony, double deceptions, and complex narratives. Instead, we focus on younger children and consider tasks designed to assess children's understanding of desires, emotions, knowledge, and beliefs. These tasks are different in focusing on different states (e.g.,

Henry M. Wellman and David Liu, Center for Human Growth and Development, University of Michigan.

Funding for this research was provided by National Institutes of Health Grant HD-22149 to Henry Wellman and by a National Science Foundation graduate fellowship to David Liu. We thank the children, parents, and the staff of the University of Michigan Children's Centers, Gretchen's House Six, and St. Joseph Mercy Hospital Child Care for their participation. We thank Shannon Duffany and Angela Kovalak for their help with data collection, and Eric Camburn for his help with Rasch model analyses.

Correspondence concerning this article should be addressed to Henry Wellman, Center for Human Growth and Development, University of Michigan, 300 N. Ingalls, 10th Floor, Ann Arbor, MI, 48109-0406. Electronic mail may be sent to hmw@umich.edu.

wants vs. thoughts). Nonetheless, these states are all similarly mental. In particular, mental states such as desires, emotions, knowledge, and beliefs can be discrepant from reality (e.g., desires vs. outcomes, actuality vs. belief) and discrepant across individuals, (e.g., when two persons have different desires for the same object or different beliefs about the same situation).

Potentially, a scaled set of tasks may have several advantages. It could more comprehensively capture children's developing understandings across a range of conceptions. A scale, based on sequences within children, provides stronger evidence for sequences than do inferences from group means. Establishing sequences of development would help constrain theorizing about theory-of-mind development. Moreover, a scaled set of tasks could provide a better measure to use in individual differences research examining the interplay between theory–of-mind understanding and other factors. This would include both the role of independent factors (e.g., family conversations, language, executive function) on theory of mind and the role of theory of mind as an independent factor contributing to other developments (e.g., social interactions, peer acceptance). Currently, research on these antecedents and consequents has been limited to simply using children's understanding of false belief as a marker of their theory-of-mind development (e.g., Astington & Jenkins, 1999; Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991; Lalonde & Chandler, 1995). However, if the intent of such studies is to index a broader construct of, and variation in, children's developing mental-state understanding, then a scale would capture such variation more informatively.

Study 1

We conducted a simple meta-analysis to summarize prior research comparing one type of theory-of-mind reasoning with another to inform our selection of scale tasks for Study 2. Most studies comparing two different theory-of-mind tasks compare performance on one sort of false-belief task (e.g., a change-of-locations task) with performance on another sort of false-belief task (e.g., an unexpected-contents task), or compare performance on a standard false-belief task with a modified false-belief task. Such comparisons were reviewed by Wellman, Cross, and Watson (2001). In Study 1 we analyzed, instead, comparisons across different mental states, for example, between children's understanding of desires versus beliefs. We aimed for a general picture of which mental-state concepts were easier than others in the preschool-age period. We did not aim for a comprehensive meta-analysis, such as Wellman et al., that closely examined moderating effects (e.g., task type or nature of protagonist).

Obviously, a pair of tasks might yield different performances either because of conceptual differences between the tasks or because of less relevant differences between task demands or features (e.g., one requiring open-ended explanations vs. the other requiring yes–no judgments). We included only pairs of tasks where the formats and demands were similar and parallel.

Method

Sample of Studies and Conditions

We began by considering all the studies listed by Wellman et al. (2001), studies that typically included the key words *belief* or *false belief* in their titles. We supplemented those studies with a computerized search of the PsycINFO database (from 1987 through 2002). We searched for studies that included the key words *desire*, *belief*, *knowledge*, *ignorance*, and *emotion* in pairwise combinations (e.g., studies whose key words included both *desire* and *belief* or *belief* and *ignorance*, and so on). We also constrained this search to include only articles focusing on children and cognition. These two sources yielded a set of more than 600 research publications for initial consideration. In addition, we scanned, more haphazardly, conference abstracts from the Society for Research in Child Development and the Cognitive Development Society. By this process we gathered as many potentially relevant studies as we could find, but we did not comprehensively search through all published and unpublished research.

From the research we examined initially, to be included in our meta-analytic comparisons a study had to provide study details in English, had to report data for preschool children, and had to report comparable data for children's performance on tasks contrasting two constructs (e.g., desire vs. belief). Moreover, the contrasting tasks in any comparison had to be closely comparable in formats, materials, and questions. Many conceivable comparisons (e.g., between understandings of desire vs. knowledge) were not represented in the literature, were represented by only one or two comparisons in only one or two studies, or employed tasks with widely varying formats and demands. Because of these limitations, as shown in Table 1, we focused on three primary comparisons. Table 1 lists the studies and conditions we used for our quantitative comparisons. The names for different conditions as listed in

Table 1
Studies and Conditions Used for the Meta-Analysis in Study 1

Study	Condition	Mean age	Mean sample size	RD
Belief vs. false belief				
Bartsch (1996)	Study 1: Discrepant belief (XX) vs. false belief (XO)	3.50	20	.53
Bartsch (1996)	Study 2: Discrepant belief (XX) vs. false belief (XY)	3.42	24	.68
Gopnik & Slaughter (1991)	Experiment 2: Image (diverse thoughts) vs. belief	3.95	24	.29
Gopnik, Slaughter, & Meltzoff (1994)	Experiment 1: Diverse belief (Level 2 think) vs. false belief	3.58	14	.64
Gopnik et al. (1994)	Experiment 2: Diverse belief (Level 2 think) vs. false belief	3.67	12	.75
Gopnik et al. (1994)	Experiment 3: Diverse belief (Level 2 think) vs. false belief	3.58	18	.30
Harris, Johnson, Hutton, Andrews, & Cooke (1989)	Experiment 3 belief (nonpreferred) vs. Experiment 2 false belief	5.34	18	.42
Wellman & Bartsch (1988)	Study 3: Not-own belief vs. explicit false belief	4.13	40	.34
Wellman & Bartsch (1988)	Study 3: Discrepant belief vs. explicit false belief	3.67	16	.66
Wellman, Hollander, & Schult (1996)	Study 1 subjective thoughts vs. Study 4 false belief	4.08	31	.18
Desire vs. belief				
Flavell, Flavell, Green, & Moses (1990)	Study 1: Value belief vs. fact belief	3.25	32	.38
Flavell et al. (1990)	Study 2: Value belief vs. fact belief	3.08	16	.69
Flavell et al. (1990)	Study 3: Value belief vs. fact belief	3.17	20	.20
Flavell et al. (1990)	Study 4: Value belief vs. fact belief	3.25	32	.33
Gopnik & Slaughter (1991)	Experiment 1: Desire vs. belief	4.00	36	.17
Gopnik & Slaughter (1991)	Experiment 2: Desire vs. belief	3.95	24	.13
Gopnik & Slaughter (1991)	Experiment 1: Intentions vs. belief	4.00	36	.17
Ruffman, Slade, & Crowe (2002)	Time 1: Desire-emotion vs. transfer (false belief)	3.01	82	.27
Ruffman et al. (2002)	Time 2: Desire-emotion vs. transfer (false belief)	3.41	79	.21
Ruffman et al. (2002)	Time 2: Desire-action vs. contents (false belief)	3.41	79	.17
Ruffman et al. (2002)	Time 3: Desire-action vs. contents (false belief)	4.04	72	.24
Wellman & Woolley (1990)	Study 2: Not-own desire vs. not-own belief	3.00	20	.20
Wellman & Woolley (1990)	Study 2: Not-own desire vs. discrepant belief	3.00	20	.60
Knowledge vs. false belief				
Fabricius & Khalil (2003)	Study 1: Know (contents) vs. false belief (contents)	5.00	84	.21
Fabricius & Khalil (2003)	Study 1: Know (location) vs. false belief (location)	5.00	84	.08
Fabricius & Khalil (2003)	Study 2: Know (contents) vs. false belief (contents)	5.33	48	.07
Fabricius & Khalil (2003)	Study 2: Know (location) vs. false belief (location)	5.33	48	−.06
Fabricius & Khalil (2003)	Study 3: Know (contents) vs. false belief (contents)	5.58	32	−.19

Continued

Table 1
Continued

Study	Condition	Mean age	Mean sample size	RD
Fabricius & Khalil (2003)	Study 3: Know (location) vs. false belief (location)	5.58	32	– .19
Flavell et al. (1990)	Study 1: Knowledge vs. fact belief	3.25	32	.05
Hogrefe, Wimmer, & Perner (1986)	Experiment 1: Ignorance vs. false belief	4.50	51	.36
Hogrefe et al. (1986)	Experiment 2: Ignorance vs. false belief	4.50	70	.14
Hogrefe et al. (1986)	Experiment 3: Ignorance vs. false belief	3.58	22	.25
Hogrefe et al. (1986)	Experiment 4: Ignorance vs. false belief	4.13	36	.27
Hogrefe et al. (1986)	Experiment 5: Ignorance vs. false belief	3.67	36	.44
Hogrefe et al. (1986)	Experiment 6: Ignorance vs. false belief	5.44	36	.35
Friedman, Griffin, Brownell, & Winner (2001)	Study 1: Ignorance (location) vs. belief (location)	4.50	54	.39
Friedman et al. (2001)	Study 1: Ignorance (contents) vs. belief (contents)	4.50	49	.31
Friedman et al. (2001)	Study 2: Ignorance vs. belief	4.50	62	.13
Roth & Leslie (1998)	Study 1: Know (false-belief task) vs. predict (false-belief task; Figure 3)	3.50	47	.44
Sullivan & Winner (1991)	Ignorance (standard) vs. false belief	3.44	44	.03
Sullivan & Winner (1991)	Ignorance (trick) vs. false belief	3.44	71	.19
Sullivan & Winner (1993)	Ignorance (standard) vs. false belief	3.55	25	– .08
Sullivan & Winner (1993)	Ignorance (trick) vs. false belief	3.55	26	– .12
Surian & Leslie (1999)	Study 2: Know vs. think (false belief)	3.33	40	.23

Note. RD = risk difference.

that table adhere as closely as possible to the names used in the original articles (with additional brief description added by us in parentheses).

Study Comparisons

Comparisons focusing on belief versus false belief in Table 1 essentially compared judgments of diverse belief versus false belief. In a diverse-belief task, truth is unknown to the child who judges that two people have differing beliefs about this (unknown) state of affairs. In false-belief tasks, in contrast, the child knows the truth. Thus, the two persons' beliefs not only differ, one person is correct and one person mistaken (i.e., has a false belief). A typical comparison is that between a not-own belief task and a false-belief task in Wellman and Bartsch (1988; see Table 1). In each task children saw a cardboard character (e.g., Bill) and a depiction of two locations (e.g., a classroom and a playground). In the not-own belief task the child was told Bill was looking for his bag, which might be in the classroom or on the playground. Then the child was asked where he or she thought the bag was likely to be. Whatever the child chose, he or she was told Bill had the opposite belief (e.g., on the playground not in the classroom) and he or she was then asked to predict what Bill would do (e.g., go to the classroom or go to the playground). Note that in such a task the child does not know where Bill's bag really is. In the comparable false-belief task (an explicit false-belief task) the child again saw a picture of Bill and of two locations (e.g., a classroom and playground). The child was told, ''Bill's bag is really on the playground,'' yet ''Bill thinks his bag is in the classroom.'' The child was then asked to predict Bill's behavior (e.g., go to the classroom or go to the playground). To be correct, the child predicts Bill's behavior on the basis of Bill's false belief not his or her own true belief. The two tasks, belief and false belief, thus use comparable materials, formats, and questions. But one targets children's understanding that beliefs can diverge between people, thus affecting behavior, and the other targets children's understanding that someone can believe something directly counter to reality, thus affecting behavior.

Comparisons focusing on desires versus beliefs were more diverse, but again each comparison listed in Table 1 used comparable tasks for the judgments compared within a study. One sort of comparison

between desires and beliefs was that between diverse beliefs and diverse desires. For example, the diverse-belief task for Wellman and Woolley (1990) was identical to the not-own belief task described earlier. Performance on that task was compared with a not-own desire task, which described two outcomes (e.g., Bill could play with puzzles in the classroom or play with sand on the playground), then asked the child's preference (e.g., play with sand), then attributed the opposite preference to the target character (e.g., Bill likes puzzles the best), and asked the child to predict Bill's action. Thus, using similar formats, one task asked the child to predict the action resulting from diverse beliefs and the other to predict the action resulting from diverse desires.

A second sort of comparison between desires and beliefs compared judgments of conflicting preferences versus conflicting beliefs, as in Flavell et al. (1990). For preferences (called *value beliefs* in that study) the child had to judge that a cookie that tastes yummy to him or her actually tastes yucky to someone else. For belief (called *fact beliefs*) the child had to judge that while he or she thinks a cup has X (which it does), someone else thinks it has Y instead.

A third sort of comparison between desire and belief concerned judgments of outdated (thus satiated) desires versus outdated (thus false) beliefs, as in Gopnik and Slaughter (1991). For desires, the child first chose one of two things (e.g., read Book A or Book B) as his or her preference, was satiated on that (e.g., read Book A), and then chose the second option as his or her current preference. The child was then asked to name his or her prior desire. For belief, the child was shown, for example, a crayon box, and then after saying he or she thought there were crayons inside, the child was shown that there really were candles inside. Then the child was asked to name his or her prior belief. Thus, the several comparisons between desires and beliefs used a variety of tasks, but in each case that we have included, the contrasting desire and belief tasks were made comparable in format and question structure.

Comparisons focusing on knowledge versus false belief in Table 1 compared judgments of ignorance versus false belief. For knowledge or ignorance judgments, the question is whether someone knows or does not know the true state of affairs. For false-belief judgments, the question is whether someone believes the true state of affairs or has a definite, alternative belief, one that contradicts reality. In a prototypical task, a character puts an object in Location A and does not see it moved to Location B. For a knowledge judgment, the child is asked if the character knows (or does not know) where the object is. For a false-belief judgment, the child is asked if the character thinks the object is in A or B. As this example shows, task materials and formats can be comparable.

Quantifying Study Comparisons

For each comparison in Table 1 we tabulated the proportion of correct responses to each contrasting task pair and the number of children (or sample size). Then, using procedures outlined in Deeks, Altman, and Bradburn (2001) and Rosenthal (1991), we calculated the risk difference (RD, or alternately *d′*), a measure of effect size indicating the size of the difference between contrasting conditions or judgments.

Results

Table 2 lists the combined results. For each set of comparisons (e.g., the 13 RD scores comparing desire vs. belief in Table 1) we list the range from highest to lowest and the mean RD. Note that RD, as calculated from these data, can be positive or negative. For example, a positive RD for desire versus belief would represent a study contrast showing desire judgments to be higher than belief judgments. A

Table 2
Meta-Analytic Comparisons and Combined Effects for Study 1

Comparison	No. of contrasts	Range	Mean RD	Random-effects-weighted mean RD	*SE*	95% CI lower bound	95% CI upper bound
Belief vs. false belief	10	.18 – .75	.48	.47	.07	.33	.61
Desire vs. belief	13	.13 – .69	.29	.29	.05	.20	.38
Knowledge vs. false belief	22	– .19 – .44	.15	.15	.04	.07	.23

Note. RD = risk difference; CI = confidence interval.

negative value would represent a contrast showing belief performance to be higher than desire.

Because diverse sets of studies (diverse in terms of the specific tasks used to measure concepts of desire, belief, or false belief across studies, and diverse in terms of the ages sampled) are grouped within each set of comparisons, the heterogeneity statistic is significant for belief versus false belief, $\chi^2(9) = 29.86$, $p < .01$; desire versus belief, $\chi^2(12) = 23.75$, $p < .05$; and knowledge versus false belief, $\chi^2(21) = 84.06$, $p < .01$. When heterogeneity is significant, a random-effects model, which incorporates both between- and within-study variance, is recommended for estimating combined effects (see Deeks et al., 2001). With a random-effects model, standard error for the combined effect increases with greater between-study variance and thus is more conservative (i.e., produces a wider confidence interval). We used the DerSimonian and Laird (1986) random-effects model to estimate combined effects, and Table 2 lists the random-effects-weighted mean RD, using the inverse variance method for combining conditions (see Deeks et al., 2001). This approach weighs each condition by the reciprocal of the within-study variance plus the between-study variance, thus taking into account sample size differences and heterogeneity and allowing for the estimation of the combined effect from a diverse set of studies.

Based on these procedures, Table 2 also shows the 95% confidence interval around each weighted mean RD. If studies show a random scatter of performance, sometimes favoring one concept but sometimes the other, the random-effects-weighted mean RD is expected to be zero. As shown in Table 2, the 95% confidence interval fails to include zero for all three contrasts. Therefore, each contrast significantly exceeds zero.

As a rule of thumb, mean RD is on the same scale as correlations and therefore can be considered small if it is in the .10 range, moderate in the .30 range, and large in the .50 range (Cohen, 1988). These data thus show moderate, but clearly significant, differences pooled across numerous studies for children's understanding of belief over false belief and desire over belief. Indeed, in these cases every RD from all studies is above zero. The results also show a smaller, yet significant, advantage for judging knowledge over false belief. In this case, although the data across studies are less consistent, with RD sometimes negative and sometimes positive, the average RD is reliably above zero.

A potential confound for estimating effect sizes might be ceiling effects that could decrease such estimates with increasing age. For example, if children typically develop concept X at 4 years of age and concept Y at 5 years of age, then examining the effect size between concepts X and Y in a sample of 6-year-olds (who would be largely at ceiling on both concepts) would underestimate any difference. However, within each of our three sets of comparisons, effect size does not correlate with mean age, all $ps > .05$. This result indicates that, given the range of ages sampled in the studies included here, potential ceiling effects do not significantly influence our estimates of effects size.

Discussion

Conceivably, all mental states might be equally hard for children to understand: All are nonobvious, internal states, and all are potentially at odds with overt behavior or external reality. Equally conceivable, children might understand some states before others, but early-understood versus late-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, the meta-analytic data show distinct regularities in children's developing understanding of mind.

The meta-analytic contrast between desires and beliefs confirms a conclusion first advanced by Wellman and Woolley (1990) and now advocated more widely (e.g., Astington, 2001; Flavell & Miller, 1998; Repacholi & Gopnik, 1997) that, on comparable tasks, children correctly judge persons' desires before they correctly judge their beliefs. The meta-analysis provides quantitative support across studies for this claim.

The comparison between belief and false belief is equally consistent and empirically stronger in the meta-analysis. These data show that children can correctly judge persons' diverse beliefs before they can judge false beliefs, a claim that has been advanced in places (e.g., Wellman et al., 2001) but not previously tested systematically across studies. Specifically, in cases where the child does not know what is true, young children can first (a) correctly judge that two persons have different beliefs, and (b) correctly judge how a person's action follows from their beliefs (in contrast to the child's own opposite belief). Only later can children correctly make the same judgments when they do know what is true and hence can (c) correctly judge that one person's belief is true and the other person's belief is decidedly false, and (d) correctly judge how a person's actions mistakenly follow from a false belief.

The data also demonstrate that children understand ignorance (e.g., that Bill does not know what is in a container) before understanding false belief (e.g., that Bill falsely believes X is in a container). This possibility, first proposed by Hogrefe et al. (1986), has been controversial. For example, Perner (1995, 2000) has argued that ignorance judgments are, necessarily, methodologically easier than false-belief judgments. He contends that young children have a default theory of belief that people look for an object where it is and that people believe what is true. Thus, baseline performance on a false-belief task with two options is 0%. In contrast, young children have no such default expectation about knowing—sometimes people know; sometimes they do not know. Thus, baseline performance on a knowledge-ignorance task with two options is 50%. Therefore, from this perspective, performance above 50% means something different in the two tasks and is always easier to achieve for an ignorance judgment versus a false-belief judgment. This baseline difference alone could account for the meta-analytic difference we found.

We are not convinced that this is the proper perspective, however. First, young children often underattribute ignorance, judging that both knowledgeable and ignorant protagonists are knowledgeable (e.g., young 3-year-olds in Woolley & Wellman, 1990, attributed ignorance 33% of the time to such protagonists rather than 50%). Yet, if Perner's (1995, 2000) argument is correct, even the youngest children's performance on ignorance judgments should average 50% correct. In the meta-analysis, the knowledge-ignorance condition with the youngest mean age was from Flavell et al. (1990). They found ignorance performance to be 36%, clearly below 50%. Thus, it is not clear empirically that young children's baseline for ignorance judgments is 50% whereas for false belief it is 0%. Second, note that some studies actually report false-belief judgments to be easier than ignorance judgments on comparable two-option tasks (i.e., studies with negative RDs in Table 1, such as Sullivan & Winner, 1993). Such a finding is difficult to square with Perner's contention. Moreover, the meta-analytic results show that the knowledge versus false belief comparison, although significant, is smaller than some others, whereas Perner's argument suggests it should be especially, artifactually, large. Therefore, a baseline difference does not adequately account for the meta-analytic difference we have found. Instead, the meta-analysis indicates that understanding ignorance develops significantly earlier than understanding false belief.

Comparisons between beliefs versus emotions exemplify a contrast that we did not include in our analyses because in all the comparisons we found the tasks were very different. To illustrate, several studies have compared children's understanding of emotions as assessed by Denham's (1986) test versus assessments of the same children's understanding of false belief (Cutting & Dunn, 1999; Hughes & Dunn, 1998; Hughes, Dunn, & White, 1998; Olson, Liu, Kerr, & Wellman, 2003). The false-belief tasks were as described earlier. In contrast, the Denham's test summed up children's performance on emotion identification items (properly labeling various emotion expressions) and on emotion attribution scenarios (e.g., attributing happiness to a character who gets ice cream). Thus, the task formats, materials, and questions used in such an emotion versus false belief comparison would be different. If we ignore those task differences, the four studies just listed yield six contrasts between false belief and emotion. If we calculate RD for these comparisons we find a mean RD of 0.41 (*range* = 0.26 to 0.53) and a random effects weighted mean RD of 0.46 (*SE* = .004). The random-effects-weighted mean RD in this case is significantly greater than zero and is in the moderate to large range. Thus, understanding of emotion as measured by the Denham test consistently precedes understanding of false belief. But, given the great differences in task formats and question structures, it is unclear how to interpret this difference.

The meta-analytic findings we present are preliminary in several senses. The relevant studies are few (providing as few as 10 contrasts for a comparison) and we are not confident we have uncovered all of the relevant published and, especially, unpublished results. Fortunately, the results are only meant to be preliminary in the additional sense of informing the design of Study 2. For this purpose the meta-analysis does show reliable differences in children's understanding of different mental states as assessed in comparable task formats. Such findings suggest that it might be possible to construct a theory-of-mind scale such that as children get older they would pass a progressively greater number of items. We tackle this possibility in Study 2.

Study 2

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 would be more valid and useful to the extent that it reflects an underlying conceptual progression or trajectory (Guttman, 1944, 1950). We reasoned that mental states such as desires, knowledge, and beliefs, albeit different in

many respects, 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, knows something but acts ignorant, or believes something not really true. Theoretically, these contrasts all reflect the fact that mental states can be said to be subjective rather than objective in varying ways. In these terms, our scale was aimed at addressing increasing steps in understanding mental subjectivity.

Based on the preliminary findings from Study 1, Study 2 includes tasks assessing diverse desires, diverse beliefs, knowledge and ignorance, and false belief. Furthermore, we reasoned that children's understanding of emotion, particularly how emotions connect with beliefs and desires, is also an important part of developing preschool theories of mind. Therefore, two other tasks involving emotion were included to capture a still broader developmental progression. One task (Belief–Emotion, as described in the Appendix) addresses how emotions connect to real situations versus to thoughts, and it is comparable in format to false-belief tasks. Another task we included (Real–Apparent Emotion, as described in the Appendix) addresses the distinction between felt versus displayed feelings.

As noted earlier, our goal was to assemble a set of tasks that are easier or harder because of conceptual differences among them (e.g., targeting desires vs. beliefs) not because of less relevant task–performance differences (e.g., one task requiring pointing, one requiring verbal judgments, one requiring written responses). Yet, strict task equivalence is often achievable only with pairs of tasks designed to compare a single conceptual contrast within a narrow age range. Thus, a consistent concern for developmental scale construction is devising tasks that span a range of ages and contents and yet are comparable or equivalent in formats and demands. We addressed this concern in several steps. We began with tasks representative of those used in the literature to connect our findings and our scale to existing studies and discussions. Each of the seven tasks included in this study was comparable to tasks in other published research, as detailed in the Appendix. However, we modified the tasks in several fashions to use more strictly comparable formats, materials, and questions across the tasks. These modifications were modest, however, to preserve the tasks' original structure and content. As a result, our tasks are not strictly comparable in all task features. Therefore, we analyzed the data in several fashions to address issues of task difficulty and comparability. One way, among others, that we addressed this issue was to include two false-belief tasks. The two false-belief tasks were not intended to yield sequentially different performance but to be roughly comparable. These two tasks could then be used to compare children's responding across different formats when the conceptual content is meant to be the same (i.e., false belief).

Method

Participants

Seventy-five 3-, 4-, and 5-year-olds (*range* = 2 years, 11 months to 6 years, 6 months) participated. Specifically, there were twenty-five 3-year-olds ($M = 3{,}7$; *range* = 2,11 to 3,11; 12 girls, 13 boys), twenty-five 4-year-olds ($M = 4{,}6$; *range* = 4,1 to 4,11; 10 girls, 15 boys), and twenty-five 5-year-olds ($M = 5{,}7$; *range* = 5,0 to 6,6; 11 girls, 14 boys). The children came from three preschools serving a population that was largely European American with approximately 25% Asian American, African American, and Hispanic American representation.

Tasks

Table 3 gives a brief description of the seven tasks, ordered in terms of their difficulty in our data (with children's percentage correct performance in parentheses). The Appendix provides a fuller description. For ease of presentation, all tasks used similar toy figurines for the target protagonists. Wellman et al. (2001) showed that, for false-belief tasks, children answer similarly when asked about real persons, videoed persons, dolls, toys, or story drawings.

Beyond using similar toy figurines, all tasks were similar 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 tasks were also similar 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

Table 3
Brief Description of Tasks in the Scale

Task	Description
Diverse Desires (95%)	Child judges that two persons (the child vs. someone else) have different desires about the same objects.
Diverse Beliefs (84%)	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 (73%)	Child sees what is in a box and judges (yes–no) the knowledge of another person who does not see what is in a box.
Contents False Belief (59%)	Child judges another person's false belief about what is in a distinctive container when child knows what it is in the container.
Explicit False Belief (57%)	Child judges how someone will search, given that person's mistaken belief.
Belief Emotion (52%)	Child judges how a person will feel, given a belief that is mistaken.
Real-Apparent Emotion (32%)	Child judges that a person can feel one thing but display a different emotion.

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.

However, to preserve the parallels between our tasks and those used in the literature certain differences remained across them. To account for these differences, in part, we chose subsets of the tasks that would be still more closely comparable in props, materials, and question formats. Specifically, Diverse Desires, Diverse Beliefs, and Explicit False Belief (Tasks 1, 2, and 5 in Table 3) formed one subset of tasks in which children saw a toy figure and a paper with two picture choices (e.g., cookie–carrot or bushes–garage). Their answer was always a verbal choice between one of these pictured choices, and the formats and questions asked were similar (as can be seen in the Appendix). Knowledge Access, Contents False Belief, and Belief–Emotion (Tasks 3, 4, and 6 in Table 3) formed a different subset in which children saw a container with a hidden item inside (e.g., a drawer with a toy dog inside, a Band-Aid box with a pig inside), and their answers were always verbal choices (e.g., "Does he think there are Band-Aids or a pig?"), although again their two response options were both embodied in the task materials. Across these three tasks, formats were similar (as can be seen in the Appendix). The Real–Apparent Emotion task, similar to other tasks, involved a toy figure, pictures (of three emotional expressions), and a short verbal story. The Real–Apparent Emotion task had a format different from any other task but was most similar to the Diverse-Desire, Diverse-Belief, and Explicit False-Belief tasks, where children made their judgment among pictured choices.

Note that each of the two primary subsets of tasks (Diverse Desires, Diverse Beliefs, and Explicit False Belief, which used toy figures and pictures, vs. Knowledge Access, Contents False Belief, and Belief–Emotion, which used toy figures and containers) included a false-belief task. As noted in our introduction, these two tasks were included, in part, to assess whether children's responding to these two different formats would be similar when conceptual content was meant to be the same. Appropriately, children performed similarly on these tasks, where 59% were correct on Contents False Belief and 57% were correct on Explicit False Belief, McNemar's $\chi^2(1) = 0$.

Procedures

Children were tested in a quiet room in their preschool by one of four adult experimenters. The seven tasks were presented in one of three orders. In all orders the Diverse-Desire 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 Real–Apparent Emotion task appeared last or next to last. Otherwise, the three orders were composed by scrambling the tasks into three different sequences; 37 children received Order 1, 19 received Order 2, and 19 received Order 3.

Results and Discussion

Table 3 shows the proportion of children correct on the various tasks, ordered from the easiest to the hardest tasks in terms of children's performance. An initial 3 (age) $\times$ 3 (order) $\times$ 2 (gender) analysis of variance (ANOVA) was conducted by giving children a score of total number correct (out of seven possible tasks). This revealed a significant main

Table 4
Guttman Scalogam Patterns for a Five-Item Scale

Pattern	1	2	3	4	5	6	Other patterns	*N*
Diverse Desire	–	+	+	+	+	+		
Diverse Belief	–	–	+	+	+	+		
Knowledge Access	–	–	–	+	+	+		
Contents False Belief	–	–	–	–	+	+		
Real-Apparent Emotion	–	–	–	–	–	+		
Participant								
3-year-olds	1	2	8	4	1	0	9	25
4-year-olds	0	2	3	2	9	5	4	25
5-year-olds	0	0	0	2	9	12	2	25
Total	1	4	11	8	19	17	15	75
Average age	3–5	4–0	3–9	4–6	4–11	5–4	4–1	

Note. A minus sign means a child failed the task in question; a plus sign means the child passed. The 6 focal patterns represent 6 of the total possible 32 patterns of response encompassing the five dichotomous items. A child exhibiting any of the remaining 26 patterns was classified as other.

effect only for age, $F(2, 57) = 25.45$, $p < .001$. With increasing age, children passed more tasks. There were no effects of task order or gender and no significant interactions. Additional analyses also confirmed that there was no significant difference in children's response to the Diverse-Desire task if they received it first or second and no difference if children received the Real–Apparent Emotion task last or next to last.

As is clear in Table 3, performance on some pairs of items was essentially equivalent (e.g., Contents False Belief and Explicit False Belief, as just mentioned). Nonetheless, as shown in the table, the tasks form a general progression. Therefore, we next examined responses to the seven tasks to see whether a subset of items formed a strict Guttman scale. This was done by initially scrutinizing the data only from the first participants tested ($n = 37$). From this examination, we found that the five items listed in Table 4 formed a reproducible Guttman scale. We then confirmed this result on participants tested last ($n = 38$); the same five items again formed a reproducible Guttman scale. Based on these initial, confirmatory analyses, we analyzed the scale properties of these items for the entire sample ($N = 75$). Table 4 shows the resulting Guttman scalogram for these five tasks.

Five-Item Guttman Scale

Guttman (1944, 1950) argued for scales where items can be ranked in difficulty such that if a person responds positively to a given item, that person must respond positively to all easier items. Thus, theoretically, a given score on a Guttman scale can only be reached with one pattern of response, and if we know a person's score, we know how that person responded to all items in the scale. Guttman scaling, or scalogram analysis, then, is the estimation of reproducibility given knowledge of person scores, that is, the extent to which item responses fit the ideal patterns. As shown in Table 4, the responses of 80% of the children (60 of 75) fit this five-item Guttman scale exactly. The coefficient of reproducibility, using Green's (1956) method of estimation, from a scalogram analysis of these data was .96 (values greater than .90 indicate scalable items). Green's index of consistency, which tests whether the observed coefficient of reproducibility was greater than what could be achieved by chance alone, was .56 (values greater than .50 are significant). Thus, these five tasks form a highly scalable set. Moreover, as children get older they tend to pass more items in succession; the relationship between age (in months) and scale score (summing the items passed out of five) is high, $r(75) = 0.64$, $p < .001$.

When children failed an item they tended to pass the relevant control questions, showing comprehension of the task formats and questions. Of course, the youngest children tended to fail both the control questions and the target questions for the very hardest items (e.g., Real–Apparent Emotion). The most relevant data, thus, concern the first task a child failed. That is, consider the tasks as ordered in Table 4. Children tended to pass the easier tasks, then reached a task where they failed, and then failed still harder tasks (a pattern that is significant in the scalogram analysis). On the first task children failed, in this order, they were 89% correct on the paired

control question and 0% correct on the target question. Thus, children largely understood the task format for the first task they failed, yet nonetheless failed on the target question. To reiterate, however, to be scored as passing a task for Table 4, children had to be correct on both the target question and its control question.

Given that our tasks were not identical in materials and formats, we next considered whether differences in task difficulty due to differences in materials, questions, and test formats might account for the progression across tasks, rather than differences in conceptual content. First, recall that the two false-belief tasks differed in materials and formats as much as any other two tasks. In spite of these differences, performances were nearly identical, as predicted on the basis of their conceptual similarity. Second, pairs of tasks within the larger sequence were closely equivalent in form. For example, Diverse Desire and Diverse Belief were chosen and constructed to be nearly identical except for the focus on desires versus beliefs, respectively. Knowledge Access and Contents False Belief were also highly similar in form (see the Appendix). Thus, it is important that additional pairwise comparisons confirmed that Diverse Desire was significantly easier than Diverse Belief, McNemar's $\chi^2(1) = 4.08$, $p < .05$; Diverse Belief was significantly easier than Contents False Belief, McNemar's $\chi^2(1) = 12.00$, $p < .001$; Knowledge Access was significantly easier than Contents False Belief, McNemar's $\chi^2(1) = 5.89$, $p < .02$; and Contents False Belief was significantly easier than Real–Apparent Emotion, McNemar's $\chi^2(1) = 13.88$, $p < .001$. These comparisons show that the larger scale captures not only a general progression but also a series of significant paired-task sequences. Furthermore, paired tasks within the scale, those that are very similar in format and task structure, confirm the more general progression across all the tasks.

Finally, consider the following concern. Perhaps the progression in Table 4 reflects baseline probabilities of being correct on the tasks rather than a conceptual progression. For example, Diverse Desire and Diverse Belief (the easiest tasks in Table 4) have a 50% probability of being correct by chance alone (i.e., correct responding is based on a single two-choice target question). However, Knowledge Access and Contents False Belief have a 25% probability of being correct by chance alone (i.e., correct responding on a two-choice target question and a two-choice control question). Of course, the discussion of Study 1 outlines some of the ways it is difficult to know definitely what the baseline rates of performance are. Nonetheless, to address this sort of concern we reanalyzed the data with a different scoring. In this alternative scoring we considered only children's responses on the target questions (ignoring the control questions). Thus, for every task there was now a single two-option response measure, meaning there was a 50% chance of being correct by random guessing alone on every task. (Note that scoring for the Real–Apparent Emotion task is also dichotomous; children's responses are incorrect if apparent emotion is equal to or less happy than the real emotion, and correct if it is more happy.) With this alternative scoring, the sequence shown in Table 4 remains, and only one child goes from exhibiting the predicted patterns to exhibiting some other pattern (and none goes in the reverse direction). Thus, with this alternative scoring, the scale shown in Table 4 captures 59 of 75 children (79%), whereas before it was 60 of 75 (80%). With this alternative scoring, the scale remains highly reproducible and significantly consistent.

This alternative scoring and analysis provide an important control. But, for future research that might use the scale, we recommend the original scoring. Individual children's understanding is better assessed, we believe, by including their performance on the control tasks as well, not simply their responses to the target questions alone.

Rasch Analyses

Guttman scales are stringent—items are scale appropriate only for fitting the exact step functions for increasing difficulty. Contemporary approaches to scale analysis have been developed, in part, to allow consideration of less strict scale progressions. Item-response theory (Bock, 1997; Embretson & Reise, 2000; Lord & Novick, 1968) consists of a family of mathematical measurement models for analyzing test or scale items. The most straightforward item-response-theory model, the Rasch measurement model, is a one-parameter logistic model for dichotomous items that estimates item difficulty and person ability levels (Rasch, 1960; Wright & Masters, 1982; Wright & Stone, 1979). The Rasch item-response-theory measurement model is often regarded as a probabilistic model for Guttman scaling (Andrich, 1985; Wilson, 1989). We analyzed our data with Rasch models to confirm and extend our Guttman scalogram analyses.

To preface the Rasch analyses, however, we believe that for cognitive development questions, Guttman scalogram analysis is an appropriate and useful analytical tool for establishing certain particularly

informative developmental sequences. Although we are aware of criticisms of Guttman scaling for its stringency in creating measurement scales (e.g., Festinger, 1947; Guilford, 1954; Nunnally & Bernstein, 1994), it is nonetheless impressive that our data fit the stringent criteria of Guttman scaling so well. This speaks to the precise sequential nature of mental state concepts children come to understand.

Both the Guttman scale and the Rasch measurement model order dichotomous items and persons on a single continuum (Andrich, 1985). The shared notion is that a person with a given ability level on a continuum will (likely) respond positively to items with difficulty levels less than that person's ability level and will (likely) respond negatively to items with difficulty levels greater than that person's ability level. However, the item-response functions for a Guttman scale are deterministic (i.e., stepwise) whereas the item-response functions for a Rasch model are probabilistic. Thus, the Guttman scale embodies a stricter measurement model than the Rasch model. For the Guttman model, if a person answers item N correctly, that person *definitely* answers item N-1 correctly. On the other hand, for the Rasch model, if a person answers item N correctly, that person *probably* answers item N-1 correctly. Rasch measurement models aim for precise estimation of items and persons on a single, interval continuum. When item difficulty exceeds person ability, the probability of a positive response is less than 0.5, relative to the difference in levels. When person ability exceeds item difficulty, the probability of positive response is greater than 0.5, relative to the difference in levels. When item difficulty equals person ability, the probability of a positive response is 0.5.

Five-item Rasch model. Data for the five items in the Guttman scale were analyzed with a Rasch model using the WINSTEPS/BIGSTEPS computer program (Linacre, 2003; Linacre & Wright, 1994). For numerical simplicity, the item difficulty and person ability measures on the linear logits scale were rescaled so that Contents False Belief (arbitrarily considered as the anchor task of the five tasks) had an item difficulty measure score of 5.0 on the linear scale. Table 5 shows the five items ordered from most difficult (highest measurement score) to least difficult (lowest measurement score). Not surprising, given the high coefficient of reproducibility of the five-item Guttman scale, the order of item difficulty is the same in the Rasch model as in the Guttman scale. However, the Rasch model allows for examination of relative distances between item difficulty scores. As shown in Table 5, although the five items are fairly evenly and widely spread, the differences (in score units) between successive items range from a low of about 1.2 to a high of more than 2.5. This is not a problem for the Rasch measurement model because it does not assume equal intervals between items; instead, it estimates the true interval between items.

Table 5 also shows summaries of item measurement scores, person measurement scores, and fit statistics. Rasch model fit statistics evaluate the notion that a person with a given ability level will likely respond positively to less difficult items and will likely respond negatively to more difficult items. Two types of fit statistics are estimated for each item and each person: infit, which is more sensitive to unexpected responses near the item or person's measurement level, and outfit, which is more sensitive to unexpected responses far from the item or

Table 5
Item and Person Measure Summary and Fit Statistics for the Five-Item Rasch Model

	Measure	Error	Standardized infit	Standardized outfit
Item difficulty summary and fit statistics				
Real-Apparent Emotion	7.73	0.46	0.1	− 0.1
Content False Belief	5.00	0.35	− 1.9	− 1.7
Knowledge Access	3.61	0.37	− 0.1	0.7
Diverse Beliefs	2.43	0.42	0.2	0.9
Diverse Desires	0.48	0.69	0.3	0.2
M	3.85	0.46	− 0.3	0.0
SD	2.44	0.12	0.8	0.9
Person ability summary and fit statistics				
M	4.66	1.65	− 0.5	− 0.2
SD	1.82	0.51	1.1	0.5

Note. Expected values for standardized infit and standardized outfit is a mean of 0 and standard deviation of 1.0; fit statistics > 2.0 indicate misfit.

person's measurement level (Linacre & Wright, 1994; Wright & Masters, 1982). Standardized infit and outfit statistics for individual items have an expected value of 0. Positive values greater than 2.0 indicate greater unpredictable variation than expected. Negative values suggest the scale is more deterministic than expected because Rasch models are probabilistic. Therefore, negative values are acceptable for our comparison with the Guttman scale because they actually indicate overfit (Bond & Fox, 2001). Therefore, we consider standardized fit statistics for individual items greater than 2.0 as indicating misfit (Wright & Masters, 1982).

As shown in Table 5, all five items' standardized infit and outfit statistics fall well short of 2.0, and mean fit statistics are near the expected value of 0. Mean standardized infit and outfit statistics for person ability, which indicate overall fit of individual persons to the scale, also fall well short of 2.0 and are near their expected value of 0. Therefore, these five items fit the Rasch model well.

Seven-item Rasch model. A problematic outcome of a Guttman scale's deterministic character is the fitting of items of similar difficulty levels on the same scale (Bond & Fox, 2001). For example, in Guttman scaling, if items J and K are very similar with item K only slightly more difficult, then permissible patterns of response are getting both items correct or getting both items wrong, and getting item J correct but item K wrong. However, the reverse pattern of getting item K correct but item J wrong (which is also likely when both items have similar difficulty levels) is not an acceptable pattern in Guttman scaling (and would greatly decrease reproducibility). As such, Guttman scales exclude items of highly similar difficulty even though those items represent similar constructs on the same scale. In our case, a five-item Guttman scale with two items excluded (Explicit False Belief and Belief–Emotion) has excellent model fit. Note that excluding two items does not mean they fail to represent the same theory-of-mind continuum as the five included items. Rather, it means that Contents False Belief, Explicit False Belief, and Belief–Emotion have similar difficulty levels and two are excluded because of the inability of strict Guttman scales to accommodate items of similar difficulty. Rasch measurement models are less problematic in fitting items of similar difficulty on the same scale. Considering a seven-item Rasch model clarifies that our seven items fit a single scale construct, while further confirming that Contents False Belief, Explicit False Belief, and Belief–Emotion have similar difficulty levels.

For this Rasch analysis, the item difficulty and person ability measures on the linear logits scale were again rescaled so that Contents False Belief has an item difficulty measure score of 5.0 on the linear scale. Table 6 shows the seven items ordered from most difficult (highest measurement score) to least difficult (lowest measurement score). Content False Belief, Explicit False Belief, and Belief–Emotion have similar difficulty levels (5.00, 5.10, and 5.49, respectively). Overall item fit (standardized infit, $M = -0.2$, $SD = 1.3$; standardized outfit, $M = 0.0$, $SD = 1.5$) and overall person fit (standardized infit, $M = -0.2$, $SD = 1.0$; standardized outfit, $M = -0.1$,

Table 6
Item and Person Measure Summary and Fit Statistics for the Seven-Item Rasch Model

	Measure	Error	Standardized infit	Standardized outfit
Item difficulty summary and fit statistics				
Real-Apparent Emotion	7.21	0.42	0.9	0.3
Belief Emotion	5.49	0.31	−1.0	−0.7
Explicit False Belief	5.10	0.36	1.9	2.6
Content False Belief	5.00	0.31	−1.9	−2.0
Knowledge Access	3.93	0.32	−1.7	−1.4
Diverse Beliefs	3.00	0.37	0.2	1.2
Diverse Desires	1.49	0.55	0.0	−0.1
M	4.46	0.38	−0.2	0.0
SD	1.71	0.08	1.3	1.5
Person ability summary and fit statistics				
M	4.96	1.15	−0.2	−0.1
SD	1.47	0.21	1.0	0.7

Note. Expected values for standardized infit and standardized outfit is a mean of 0 and standard deviation of 1.0; fit statistics > 2.0 indicate misfit.

$SD = 0.7$) are excellent. One item among the seven, however, has poorer fit, although not extremely poor; Explicit False Belief has a standardized infit of 1.9 and a standardized outfit of 2.6. This does not indicate that the Explicit-False-Belief item assesses a different conceptual content from the other items. Rather, this finding again demonstrates how items with similar levels of difficulty can result in poor fit—slightly in Rasch models but drastically in Guttman scaling (Andrich, 1985).

Scoring Individuals' Performance

One potential advantage of Rasch measurement models over Guttman scaling is the precision with which individual person ability scores on an interval scale can be estimated. However, scoring of individual person ability with a Guttman scale is more practical because all it involves is simply adding up the number of items answered correctly. For our data, the five-item Rasch scores and the five-item Guttman scores are almost perfectly correlated, $r(75) = .998$, $p < .001$. Furthermore, the relation between the five-item Rasch scores and age, $r(75) = .645$, $p < .001$, and the relation between the five-item Guttman scores and age, $r(75) = .638$, $p < .001$, are almost identical. Therefore, for our five-item data, person ability scores estimated with both measurement models are so similar that any precision gained with the Rasch model is outweighed by the practicality of scoring the five-item Guttman scale.

General Discussion

> The chronological order in which cognitive novelties emerge during childhood is a datum of central importance for the student of human cognitive growth. (Flavell 1972, p. 281)

The data from Study 2 demonstrate an extended series of conceptual insights that take place in the preschool years as children acquire a theory of mind. In this regard they confirm but go beyond earlier studies, for example, those encompassed in the meta-analysis of Study 1.

Empirically, the findings demonstrate an understanding of desires that precedes an understanding of beliefs; in particular, children become aware that two persons can have different desires for the same object before they become aware that two persons can have different beliefs about the same object. They also demonstrate an understanding of diverse beliefs before false beliefs; that is, children can judge that they and someone else can have differing beliefs about the same situation (when the child does not know which belief is true and which is false) before they judge that someone else can have a false belief about a situation (where the child thus knows which belief is true and which is false). Finally, the results show that differentiating between real and apparent emotion is a late-developing understanding within the preschool years.

Using Flavell's (1972) taxonomy of developmental sequences, it is clear that the sequence charted in Study 2 is not one of addition (it is not the case that understandings tapped by later items are equal alternatives to those appearing earlier in this sequence) and not one of substitution (it is not the case that later understandings replace earlier understandings; earlier understandings in this sequence remain valid and older children pass later items and earlier items as well). Instead, the sequence represents one of modification or mediation. For modification, according to Flavell, earlier items represent initial insights that are broadened or generalized to encompass later insights. In this regard, our reasoning in choosing tasks was that the tasks similarly address issues of subjectivity but encompass subjective–objective distinctions of purposefully varying sorts. For example, some items focus on subjective–subjective individuation (where two persons could have contrasting mental states about the same situation), some items focus on subjective–objective contrasts (where some situation might be objectively true, but a person is ignorant of it or mistaken about it), and some items focus on internal–external contrasts (where an initial, subjective state might be of one sort but its external, overt expression is of a different sort). That the tasks scale into a single continuum is consistent with an interpretation that children's understanding of subjectivity is progressively broadening and developing in the preschool years.

Mediation sequences go further in claiming that the earlier insights enable or aid in the attainment of later insights. In our case, it is possible to theorize that an initial understanding of the subjectivity of desires, once achieved, could mediate an understanding of the subjectivity of representational mental states such as belief. Furthermore, an understanding that two persons can have diverse beliefs in a situation where truth is not known (and so the contrast is only between two individuals' mental states), once achieved, could scaffold a later understanding of ignorance or false belief (and so the contrast is between individuals' mental states and

reality). We favor this constructivist, theoretical interpretation, but data about sequences of acquisition alone do not provide definitive support for such a strong interpretation.

At the same time, taken together, the findings from Studies 1 and 2 shed light on some contrasting theoretical claims. In particular, the progression from desire to diverse belief to false belief is of interest. Both modular accounts (e.g., Leslie, 1994) and simulation accounts (e.g., Harris, 1992) claim that preschool children equally understand beliefs and desires; it is false belief that is peculiarly and distinctively difficult. In contrast, the data from Studies 1 and 2 show that understanding beliefs is more difficult than desires, even when understanding false belief is not at issue. In Study 2, for example, the Diverse-Belief task did not require understanding false belief but was nonetheless significantly more difficult than understanding Diverse Desires in spite of being almost identical in format, materials, and so on. Alternatively, executive function accounts (or more precisely what Carlson & Moses, 2001, called *executive function expression accounts*) suggest that children's difficulty with mental states in general, and false belief in particular, stem from difficulties in inhibiting a prepotent response to generate a different response. For example, responding correctly to a false–belief task requires not stating what one knows is true but stating instead what the other person thinks is true. However, both the Diverse-Desires and Diverse-Beliefs tasks in Study 2, and the Desire versus Belief comparisons in Study 1, are similar in requiring inhibition of one's own point of view to answer in terms of the other person's point of view. Performance in belief tasks is nonetheless still worse than performance in desire tasks.

The biggest contribution of our research, however, is more descriptive than explanatory. In particular, the studies confirm that theory-of-mind understandings represent an extended and progressive set of conceptual acquisitions. No single type of task—for example, false-belief tasks—can adequately capture this developmental progression. Similarly, no theory will be adequate that does not account for these various, developmentally sequenced acquisitions. Practically, this conclusion carries the implication that a theory-of-mind scale is needed to more adequately capture individual children's theory-of-mind developments.

In this vein, our findings in Study 2 provide a battery of items that constitutes a consistent scale that captures children's developmental progression. We believe this scale has several advantages for future research on theory of mind. For example, consider again research examining the interplay between theory-of-mind understanding and other factors. This includes both the role of independent factors (e.g., family conversations, language, executive function) on theory of mind and the role of theory of mind as a factor contributing to other developments (e.g., social interactions, peer acceptance). The burgeoning research on these issues faces measurement limitations by typically using single tasks, essentially false-belief tasks, to assess children's understanding (e.g., Astington & Jenkins, 1999; Dunn et al., 1991; Lalonde & Chandler, 1995). The current scale is usable with a wider range of ages, provides a more continuous variable for comparing individuals, and captures a greater variety of conceptual content. Wellman, Phillips, Dunphy-Lelii, and LaLonde (in press) provide an initial demonstration of the scale's utility in capturing individual differences.

As another example, consider research with individuals with autism, who are significantly impaired at theory-of-mind understandings (e.g., Baron-Cohen, 1995). Significantly, high-functioning individuals with autism typically fail false-belief tasks whereas comparable normal and mentally retarded individuals pass such tasks. Yet, about 20% to 25% of high-functioning individuals with autism pass false-belief tasks (Baron-Cohen, 1995; Happe, 1994). These data raise several questions. In particular: Are individuals with autism distinctively impaired in theory-of-mind understandings or only significantly delayed? More precisely, to the extent older children with autism achieve social cognitive understandings (e.g., understanding false beliefs), does this represent delay in a consistent developmental trajectory or an ad hoc or alternatively based understanding achieved via nonordinary strategies and mechanisms? Longitudinal data from individuals with autism on a variety of tasks could address such questions. But a theory-of-mind scale, such as the present one, could also provide critical data. It could disclose whether individuals with autism who pass (or fail) false-belief tasks do or do not exhibit the normally developing progression of related understandings evident in Table 4. A theory-of-mind scale could be used to address similar, comparative questions with other populations (e.g., deaf children; Peterson & Siegal, 1995).

The current scale also has several features that could prove useful in future research. The five-item version is highly scalable (approximating a strict Guttman scale), includes a false-belief task, yet spans a larger range of ages and tasks, yielding scale scores ranging from 0 to 5. The five task items can be

administered in 15 to 20 min. Using six or seven items would include an increased array of task items, useful for more extended theoretical comparisons, but would still require only about 20 min to administer. Further information about materials and procedures are available on request from the authors.

In conclusion, the theory-of-mind scale validated in Study 2 establishes both (a) a progression of conceptual achievements that mark social cognitive understanding in normally developing preschool children and (b) a method for measuring that development accurately and informatively.

Appendix

Diverse Desires

Children see a toy figure of an adult and a sheet of paper with a carrot and a cookie drawn on it. ''Here's Mr. Jones. It's snack time, so, Mr. Jones wants a snack to eat. Here are two different snacks: a carrot and a cookie. Which snack would you like best? Would you like a carrot or a cookie best?'' This is the *own-desire* question.

If the child chooses the carrot: ''Well, that's a good choice, but Mr. Jones really likes cookies. He doesn't like carrots. What he likes best are cookies.'' (Or, if the child chooses the cookie, he or she is told Mr. Jones likes carrots.) Then the child is asked the *target* question: ''So, now it's time to eat. Mr. Jones can only choose one snack, just one. Which snack will Mr. Jones choose? A carrot or a cookie?''

To be scored as correct, or to pass this task, the child must answer the *target* question opposite from his or her answer to the *own-desire* question.

This task was derived from those used by Wellman and Woolley (1990) and Repacholi and Gopnik (1997).

Diverse Beliefs

Children see a toy figure of a girl and a sheet of paper with bushes and a garage drawn on it. ''Here's Linda. Linda wants to find her cat. Her cat might be hiding in the bushes or it might be hiding in the garage. Where do you think the cat is? In the bushes or in the garage?'' This is the *own-belief* question.

If the child chooses the bushes: ''Well, that's a good idea, but Linda thinks her cat is in the garage. She thinks her cat is in the garage.'' (Or, if the child chooses the garage, he or she is told Linda thinks her cat is in the bushes.) Then the child is asked the *target* question: ''So where will Linda look for her cat? In the bushes or in the garage?''

To be correct the child must answer the *target* question opposite from his or her answer to the *own-belief* question.

This task was derived from those used by Wellman and Bartsch (1989) and Wellman et al. (1996).

Knowledge Access

Children see a nondescript plastic box with a drawer containing a small plastic toy dog inside the closed drawer. ''Here's a drawer. What do you think is inside the drawer?'' (The child can give any answer he or she likes or indicate that he or she does not know). Next, the drawer is opened and the child is shown the content of the drawer: ''Let's see … it's really a dog inside!'' Close the drawer: ''Okay, what is in the drawer?''

Then a toy figure of a girl is produced: ''Polly has never ever seen inside this drawer. Now here comes Polly. So, does Polly know what is in the drawer? (the *target* question) ''Did Polly see inside this drawer?'' (the *memory* question).

To be correct the child must answer the *target* question ''no'' and answer the *memory* control question ''no.''

This task was derived from those used by Pratt and Bryant (1990) and Pillow (1989), although it was modified so that the format was more parallel to the contents False-Belief task.

Contents False Belief

The child sees a clearly identifiable Band-Aid box with a plastic toy pig inside the closed Band-Aid box. ''Here's a Band-Aid box. What do you think is inside the Band-Aid box?'' Next, the Band-Aid box is opened: ''Let's see … it's really a pig inside!'' The Band-Aid box is closed: ''Okay, what is in the Band-Aid box?''

Then a toy figure of a boy is produced: ''Peter has never ever seen inside this Band-Aid box. Now here comes Peter. So, what does Peter think is in the box? Band-Aids or a pig? (the *target* question) ''Did Peter see inside this box?'' (the *memory* question).

To be correct the child must answer the *target* question ''Band-Aids'' and answer the *memory* question ''no.''

This task was derived from one used initially by Perner, Leekam, and Wimmer (1987) and widely modified and used since then (see Wellman et al., 2001).

Explicit False Belief

Children see a toy figure of a boy and a sheet of paper with a backpack and a closet drawn on it. ''Here's Scott. Scott wants to find his mittens. His mittens might be in his backpack or they might be in the closet. *Really*, Scott's mittens are in his backpack. But Scott *thinks* his mittens are in the closet.''

''So, where will Scott look for his mittens? In his backpack or in the closet?'' (the *target* question) ''Where are Scott's mittens really? In his backpack or in the closet?'' (the *reality* question).

To be correct the child must answer the *target* question ''closet'' and answer the *reality* question ''backpack.''

This task was derived from one used by Wellman and Bartsch (1989) and Siegal and Beattie (1991).

Belief–Emotion

Children see a toy figure of a boy and a clearly identifiable individual-size Cheerios box with rocks inside the closed box. ''Here is a Cheerios box and here is Teddy. What do you think is inside the Cheerios box?'' (Cheerios) Then the adult makes Teddy speak: ''Teddy says, 'Oh good, because I love Cheerios. Cheerios are my favorite snack. Now I'll go play.''' Teddy is then put away and out of sight.

Next, the Cheerios box is opened and the contents are shown to the child: ''Let's see … there are really rocks inside and no Cheerios! There's nothing but rocks.'' The Cheerios box is closed: ''Okay, what is Teddy's favorite snack?'' (Cheerios).

Then Teddy comes back: ''Teddy has never ever seen inside this box. Now here comes Teddy. Teddy's back and it's snack time. Let's give Teddy this box. So, how does Teddy feel when he gets this box? Happy or sad?'' (the *target* question) The adult opens the Cheerios box and lets the toy figure look inside: ''How does Teddy feel after he looks inside the box? Happy or sad?'' (the *emotion-control* question).

To be correct, the child must answer the *target* question ''happy'' and answer the *emotion-control* question ''sad.''

This task was derived from one used by Harris, Johnson, Hutton, Andrews, and Cooke (1989).

Real–Apparent Emotion

Initially, children see a sheet of paper with three faces drawn on it—a happy, a neutral, and a sad face—to check that the child knows these emotional expressions. Then that paper is put aside, and the task begins with the child being shown a cardboard cutout figure of a boy drawn from the back so that the boy's facial expression cannot be seen. ''This story is about a boy. I'm going to ask you about how the boy really feels inside and how he looks on his face. He might really feel one way inside but look a different way on his face. Or, he might really feel the same way inside as he looks on his face. I want you to tell me how he really feels inside and how he looks on his face.''

''This story is about Matt. Matt's friends were playing together and telling jokes. One of the older children, Rosie, told a mean joke about Matt and everyone laughed. Everyone thought it was very funny, but *not* Matt. But, Matt didn't want the other children to see how he felt about the joke, because they would call him a baby. So, Matt tried to *hide how he felt*.'' Then the child gets two memory checks: ''What did the other children do when Rosie told a mean joke about Matt?'' (Laughed or thought it was funny.) ''In the story, what would the other children do if they knew how Matt felt?'' (Call Matt a baby or tease him.)

Pointing to the three emotion pictures: ''So, how did Matt really feel, when everyone laughed? Did he feel happy, sad, or okay?'' (the *target-feel* question) ''How did Matt try to look on his face, when everyone laughed? Did he look happy, sad, or okay? (the *target-look* question).

To be correct the child's answer to the *target-feel* question must be more negative than his or her answer to the *target-look* question (i.e., sad for target-feel and happy or okay for target-look, or okay for target-feel and happy for target-look).

This task was derived from one used by Harris, Donnelly, Guz, and Pitt-Watson (1986).

References

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. *Sociological Methodology, 15*, 33–80.

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

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental Psychology, 35*, 1311–1320.

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

Bartsch, K. (1996). Between desires and beliefs: Young children's action predictions. *Child Development*, *67*, 1671–1685.

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

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*, 21–33.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in human sciences*. Mahwah, NJ: Erlbaum.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.), Hillsdale, NJ: Erlbaum.

Cutting, A. L., & Dunn, J. (1999). Theory of mind, emotion understanding, language, and family background. *Child Development*, *70*, 853–865.

Deeks, J. J., Altman, D. G., & Bradburn, M. J. (2001). Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In M. Egger, D. G. Smith, & D. G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (pp. 285–312). London: BMJ Books.

Denham, S. A. (1986). Social cognition, prosocial behavior and emotion in preschoolers. *Child Development*, *57*, 194–201.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Dunn, J., Brown, J., Slomkowski, C., Tesla, C., & Youngblade, L. (1991). Young children's understanding of other people's feelings and beliefs: Individual differences and their antecedents. *Child Development*, *62*, 1352–1366.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fabricius, W. V., & Khalil, S. L. (2003). False beliefs or false positives? *Journal of Cognition and Development*, *4*, 239–262.

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

Flavell, J. H. (1972). An analysis of cognitive-developmental sequences. *Genetic Psychology Monographs*, *86*, 279–350.

Flavell, J. H., Flavell, E. R., Green, F. L., & Moses, L. J. (1990). Young children's understanding of fact beliefs versus value beliefs. *Child Development*, *61*, 915–928.

Flavell, J. H., & Miller, P. H. (1998). Social cognition. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology: Vol. 2: Cognition, perception, and language* (pp. 851–898). New York: Wiley.

Friedman, O., Griffin, R., Brownell, H., & Winner, E. (2001). *Problems with the seeing = knowing rule.* Manuscript submitted for publication.

Gopnik, A., & Slaughter, V. (1991). Young children's understanding of changes in their mental states. *Child Development*, *62*, 98–110.

Gopnik, A., Slaughter, V., & Meltzoff, A. (1994). Changing your views: How understanding visual perception can lead to a new theory of mind. In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind: Origins and development* (pp. 157–181). Hove, England: Erlbaum.

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

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

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

Guttman, L. (1950). The basis of scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. A. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Happe, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism & Developmental Disorders*, *24*, 129–154.

Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind & Language*, *7*, 120–144.

Harris, P. L., Donnelly, K., Guz, G. R., & Pitt-Watson, R. (1986). Children's understanding of the distinction between real and apparent emotion. *Child Development*, *57*, 895–909.

Harris, P. L., Johnson, C. N., Hutton, D., Andrews, G., & Cooke, T. (1989). Young children's theory of mind and emotion. *Cognition & Emotion*, *3*, 379–400.

Hogrefe, G. J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, *57*, 567–582.

Hughes, C., & Dunn, J. (1998). Understanding and emotion: Longitudinal associations with mental-state talk between young friends. *Developmental Psychology, 34*, 1026–1037.

Hughes, C., Dunn, J., & White, A. (1998). Trick or treat? Uneven understanding of mind and emotion and executive dysfunction in "hard to manage" preschoolers. *Journal of Child Psychology & Psychiatry, 39*, 981–994.

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.

Leslie, A. M. (1994). ToMM, ToBy, and agency: Core architecture and domain specificity in cognition and culture. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). New York: Cambridge University Press.

Linacre, J. M. (2003). *User's guide and program manual to WINSTEPS: Rasch model computer programs*. Chicago: MESA Press.

Linacre, J. M., & Wright, B. D. (1994). *A user's guide to BIGSTEPS: Rasch model computer programs*. Chicago: MESA Press.

Lord, F. N., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mitchell, P. (1996). *Acquiring a conception of mind: A review of psychological research and theory.* Hove, England: Psychology Press.

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

Olson, S. L., Liu, D., Kerr, D. C., & Wellman, H. M. (2003). *Social and behavioral outcomes of children's theory of mind development*. Unpublished manuscript.

Perner, J. (1995). The many faces of belief. *Cognition*, *57*, 241–269.

Perner, J. (2000). About + belief + counterfactual. In P. Mitchell & K. J. Riggs (Eds.), *Children's reasoning and the mind* (pp. 367–401). Hove, England: Taylor & Frances.

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

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.

Peterson, C. C., & Siegal, M. (1995). Deafness, conversation and theory of mind. *Journal of Child Psychology and Psychiatry*, *36*, 459–474.

Pillow, B. H. (1989). Early understanding of perception as a source of knowledge. *Journal of Experimental Child Psychology*, *47*, 116–129.

Pratt, C., & Bryant, P. E. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development*, *61*, 973–982.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*, 12–21.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Roth, D., & Leslie, A. (1998). Solving belief problems: Toward a task analysis. *Cognition*, *66*, 1–31.

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.

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

Sullivan, K., & Winner, E. (1991). When 3-year-olds understand ignorance, false belief and representational change. *British Journal of Developmental Psychology*, *9*, 159–171.

Sullivan, K., & Winner, E. (1993). Three-year-old's understanding of mental states: The influence of trickery. *Journal of Experimental Child Psychology*, *56*, 135–148.

Surian, L., & Leslie, A. M. (1999). Competence and performance in false belief understanding. *British Journal of Developmental Psychology*, *17*, 141–155.

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, England: Blackwell.

Wellman, H. M., & Bartsch, K. (1988). Young children's reasoning about beliefs. *Cognition*, *30*, 239–277.

Wellman, H. M., & Bartsch, K. (1989). 3-year-olds understand belief. *Cognition*, *33*, 321–326.

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., & Hickling, A. K. (1994). The minds "I": Children's conception of the mind as an active agent. *Child Development*, *65*, 1564–1580.

Wellman, H. M., Hollander, M., & Schult, C. A. (1996). Young children's understanding of thought-bubbles and of thoughts. *Child Development*, *67*, 768–788.

Wellman, H. M., Pillips, A. T., Dunphy-Lelii, S., & LaLonde, N. (in press). Infant social attention predicts preschool social cognition. *Developmental Science*.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*, 245–275.

Wilson, M. (1989). A comparison of deterministic and probabilistic approaches to measuring leaning structures. *Australian Journal of Education*, *33*, 127–140.

Woolley, J., & Wellman, H. M. (1993). Origin and truth: Young children's understanding of imaginary mental representations. *Child Development*, *64*, 1–17.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.