

Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief

Henry M. Wellman, David Cross, and Julianne Watson

Research on theory of mind increasingly encompasses apparently contradictory findings. In particular, in initial studies, older preschoolers consistently passed false-belief tasks—a so-called “definitive” test of mental-state understanding—whereas younger children systematically erred. More recent studies, however, have found evidence of false-belief understanding in 3-year-olds or have demonstrated conditions that improve children’s performance. A meta-analysis was conducted ($N = 178$ separate studies) to address the empirical inconsistencies and theoretical controversies. When organized into a systematic set of factors that vary across studies, false-belief results cluster systematically with the exception of only a few outliers. A combined model that included age, country of origin, and four task factors (e.g., whether the task objects were transformed in order to deceive the protagonist or not) yielded a multiple R of .74 and an R^2 of .55; thus, the model accounts for 55% of the variance in false-belief performance. Moreover, false-belief performance showed a consistent developmental pattern, even across various countries and various task manipulations: preschoolers went from below-chance performance to above-chance performance. The findings are inconsistent with early competence proposals that claim that developmental changes are due to tasks artifacts, and thus disappear in simpler, revised false-belief tasks; and are, instead, consistent with theoretical accounts that propose that understanding of belief, and, relatedly, understanding of mind, exhibit genuine conceptual change in the preschool years.

INTRODUCTION

“Theory of mind” has become an important theoretical construct and the topic of considerable research effort. Theory of mind describes one approach to a larger topic: everyday or folk psychology—the construal of persons as psychological beings, interactors, and selves. The phrase, theory of mind, emphasizes that everyday psychology involves seeing oneself and others in terms of mental states—the desires, emotions, beliefs, intentions, and other inner experiences that result in and are manifested in human action. Furthermore, everyday understanding of people in these terms is thought to have a notable coherence. Because actors have certain desires and relevant beliefs, they engage in intentional acts, the success and failure of which result in various emotional reactions. Whether or in what sense everyday psychology is theorylike is a matter of contention (see, e.g., Gopnik & Wellman, 1994; Nelson, Plesa, & Henseler, 1998). Regardless, the phrase, theory of mind, highlights two essential features of everyday psychology: its coherence and mentalism.

How, when, and in what manner does an everyday theory of mind arise? This question has generated much current research with children. The question has been investigated using a variety of tasks and studies that focus on various conceptions within the child’s developing understanding, for example, conceptions of desires, emotions, beliefs, belief–desire

reasoning, or psychological explanation, among others (see, e.g., Astington, 1993; Flavell & Miller, 1998; Wellman, 1990). From the earliest research, however, a central focus has been on children’s understanding of belief, especially false belief. Why? Mental-state understanding requires realizing that such states may reflect reality and may be manifest in overt behavior, but are nonetheless internal and mental, and thus distinct from real-world events, situations, or behaviors. A child’s understanding that a person has a false belief—one whose content contradicts reality—provides compelling evidence for appreciating this distinction between mind and world (see, e.g., Dennett, 1979).

A now classic false-belief task presents a child with the following scenario (Wimmer & Perner, 1983): Maxi puts his chocolate in the kitchen cupboard and leaves the room to play. While he is away (and cannot see) his mother moves the chocolate from the cupboard to a drawer. Maxi returns. Where will he look for his chocolate, in the drawer or in the cupboard? Four- and 5-year-olds often pass such tasks, judging that Maxi will search in the cupboard although the chocolate really is in the drawer. These correct answers provide evidence that the child knows that Maxi’s actions depend on his beliefs rather than simply the real situation itself, because belief and reality

diverge. Many younger children, typically 3-year-olds, fail such tasks. Instead of answering randomly, younger children often make a specific false-belief error—they assert that Maxi will look for the chocolate in the drawer to which it was moved.

These findings are empirically intriguing; it is striking that young children would make such a provocative error. They are also theoretically important in that they support a developmental hypothesis that children's theory of mind undergoes a major conceptual change in early life. To be clear, the claim is not that young children know nothing of mental states, but that they fail to understand representational mental states. One way of depicting this claim is presented in Figure 1. A simple understanding of a state such as desire could involve construing the desirer as having an internal, subjective urging for an external state of affairs. But an everyday understanding of belief requires the notion that the person has a representation of the world, the contents of which could be, and in the case of false beliefs are, quite different from the contents of the world itself. Thus, the change involved has been characterized as a shift (1) from a situation-based to a representation-based understanding of behavior (Perner, 1991), (2) from a connections to a representational understanding of mind (Flavell, 1988), or (3) from a simple desire to a belief-desire naive psychology (Wellman, 1990). In short, as revealed in part by successful false-belief reasoning, the claim

is that older children understand that people live their lives in a mental world as much as in a world of real situations and occurrences.

False-belief performance has come to serve as a marker for mentalistic understanding of persons more generally. Thus, in research on individual differences in young children's social cognition, false-belief performances are used as a major outcome measure to assess the influence of early family conversations (Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991), engagement in pretend play (Youngblade & Dunn, 1995), or family structure (Hughes & Dunn, 1998; Perner, Ruffman, & Leekam, 1994) on development of mentalistic understandings. False-belief understanding has also become a major tool for research with developmentally delayed individuals. The theory-of-mind hypothesis for autism, in particular, claims that the severe social disconnectedness evident in even high-functioning individuals with autism is due to an impairment in their ability to construe persons in terms of their inner mental lives (see, e.g., Baron-Cohen, 1995). False-belief performances provided an initial empirical test of this claim in that high-functioning children with autism who are able to reason competently about physical phenomena often fail false-belief tasks, whereas Down syndrome and other delayed populations of equivalent mental age often do not (e.g., Baron-Cohen, Leslie, & Frith, 1985; for more comprehensive findings and comparisons, see Happe, 1995; Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998).

For various reasons, therefore, a considerable body of research has accumulated, which employ an increasing variety of false-belief tasks that focus on attempting to demonstrate and explain false-belief errors, as well as relate performance on false-belief tasks to other conceptions, tasks, and competencies. Theory-of-mind research goes well beyond this task and these data; nonetheless false-belief tasks have a central place in current social-cognitive research (see Flavell & Miller, 1998), much as conservation tasks once were focal for understanding cognitive development and for testing Piaget's findings and theorizing. For the case of false belief, just as in the conservation literature, the initial accounts, the initial tasks, and especially the claims of conceptual change have all been vigorously challenged.

In particular, research on false belief instantiates a basic conundrum in the study of cognitive development. Performance on any cognitive task reflects at least two factors: conceptual understanding required to solve the problem ("competence") and other non-focal cognitive skills (e.g., ability to remember the key information, focus attention, comprehend, and an-

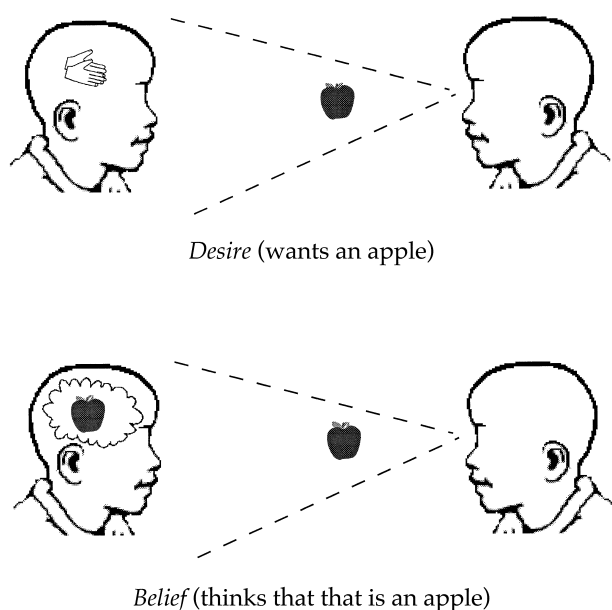


Figure 1 Graphic depiction of how the person on the right construes the desire (top) or belief (bottom) of the person on the left.

swer various questions) required to access and express understanding ("performance"). The last 25 years of cognitive development research have produced a plethora of early competence studies and accounts essentially showing that on various tasks young children fail not because they lack the conceptual competence, but rather because the testing situation was too demanding or confusing. This research has had several desirable results: undeniable underestimations of young children's knowledge have been exposed; information-processing analyses of how children arrive at their answers and responses, and not just what answers or responses they make, have flourished; and domain-general accounts of cognitive competence have yielded to more precise domain-specific understandings of children's conceptions and skills. At the same time, however, accepted demonstrations of conceptual change have largely disappeared. This is curious inasmuch as the interplay between cognitive change and stability is the cornerstone of all major theories of cognitive development. Yet each proposed developmental change (e.g., Piaget's conservation competence, Carey's proposed shift from naive psychology to naive biology, false-belief understandings) has seemingly evaporated in the mist of task variations showing enhanced performance in still younger children.

Conceivably, conceptual change may indeed be rare. The contemporary re-emergence of strongly nativist perspectives on cognitive development both contributes to and derives from this possibility (e.g., Spelke, 1994). On the other hand, genuine conceptual changes may be obscured by current emphasis on early competence and task simplifications, making it difficult to comprehend the bigger picture amidst the haze of accumulating results from numerous task variations. A comprehensive analysis of the voluminous and varied false-belief research provides an important contemporary opportunity to examine this basic issue.

Empirically, an increasing number of researchers now claim that the original false-belief tasks are unnecessarily difficult and that 3-year-olds can evidence improved, even above-chance, false-belief reasoning if the tasks are suitably revised (e.g., Siegal & Beattie, 1991; Sullivan & Winner, 1993). Not only the correct estimation of 3-year-olds is at issue, but more importantly, basic developmental trends. Some authors now claim that 3-year-olds, and much younger children as well, understand belief and false belief (e.g., Chandler, Fritz, & Hala, 1989; Fodor, 1992). False-belief competence, assessed correctly, is thus predicted to be high even in young children. Other authors (e.g., Robinson & Mitchell, 1995) claim that many 3-year-

olds fail false-belief tasks but claim many 4- and 5-year-olds do as well. "The most striking thing about the age trends was the lack of them . . . Quite simply, it has become fashionable to claim that there is a sharp age trend, but in fact there is not" (Mitchell, 1996, pp. 137–138).

These issues, controversies, and theories, along with the increasing amount of empirical findings, mandate a careful review. Indeed, several qualitative reviews have been presented, both several years ago when the number of focal studies were modest (Flavell, 1988; Perner, 1991; Wellman, 1990) and more recently as the studies and discrepancies have grown (Leslie, 2000; Mitchell, 1996; Taylor, 1996). But these reviews have come to different conclusions, and have failed to provide a compelling synthesis of all the data. We provide a quantitative review and integration of the findings, a meta-analysis of the data. A meta-analysis seems feasible and useful. The number of studies is now large; certainly sufficient for informative meta-analysis and, indeed, so large that considering all the studies one by one is difficult or impossible. In addition, although different qualitative reviews depict this database in contradictory fashions, many of the contradictions could be addressed by meta-analysis. Consider the essential dependent variable, false-belief performance as reported across studies. When studies show a mixture of above-, at-, and below-chance responding, as is the case for research on false belief, then a comprehensive pooling of the data across studies is needed to clarify the nature of children's performance.

Of equal theoretical importance are possible independent variables, such as the ages of the children tested. In particular, investigations of false belief reasoning have increasingly differed according to a variety of task variables. The focal false-belief question has been asked in terms of action (where will Maxi *look* for his chocolate), in terms of thoughts (where does Maxi *think* his chocolate is), and in terms of speech (where will Maxi *say* his chocolate is). Sometimes the target object has been moved from its original location inadvertently, whereas at other times, the movement has been presented, emphatically, as a deliberate deception designed to fool or trick the protagonist. At various times the protagonist, Maxi, has been a puppet, a doll, a real person, or a video portrayal. Moreover, the change-of-locations task discussed earlier represents only one "standard" false-belief task among several. Another often-used task involves unexpected contents: Children see a crayon box, state it will have crayons inside, then open the box to find it is filled with candies. Then they are asked about someone else, Mary, who has never

looked inside the box, "What will Mary think is in here, crayons or candies?" A variation on unexpected-contents tasks are unexpected-identity tasks, in which a false belief is engendered by an object with a deceptive identity (e.g., a sponge that looks just like a rock).

The variety of independent variables contributes to an increasingly confused picture. Again meta-analysis could be especially informative; sufficient variation exists on a number of independent variables across studies to allow for the examination of factors that have not been varied, or not comprehensively varied, within studies.

Among the extended set of variables we will consider in this study, five deserve brief introduction because of their prominence in research and theory. The first variable is age. According to initial findings, false-belief performance changes dramatically from 3 to 5 years of age; this change may or may not remain when important task variations are accounted for. The second variable concerns deception—whether the child views the false-belief situation as happenstance or a deliberate ploy to trick the protagonist. Several authors claim that framing the task in terms of explicit trickery reduces or eliminates young children's errors (e.g., Chandler et al., 1989), although others claim it does not (e.g., Sodian, Taylor, Harris, & Perner, 1991). The third variable concerns salience. For example, framing the task in terms of trickery may be unimportant except that it serves to make mental state (being fooled or duped) more salient to the child in comparison with the alternative real state of affairs (where the chocolate really is). Salience issues have two separate aspects. Certain task manipulations arguably serve to enhance the salience of the protagonist's mental state; other manipulations serve to reduce the salience of the contrasting real state of affairs.

The fourth variable of special import concerns whether the children are asked about someone else's false belief or their own. In a false-belief task for self (sometimes called a representational change task), a child may be shown the crayon box, and that it contains candies, and then asked, "Before you looked inside, did you think the box contained crayons or candies?" Different theoretical accounts make different predictions about the difficulty of false-belief judgments for self versus others (as detailed later in the Discussion section). Moreover, an empirically intriguing question is, do children understand beliefs first for their own case, for others, or both similarly?

Finally, the national or community identity of the children tested needs to be considered, for example, whether children are from the United States, Austria, Japan, or a hunter-gatherer traditional African soci-

ety. Understanding people in terms of mental states in general, and in terms of beliefs, more specifically, can be argued to be a universal folk psychological stance (e.g., Wellman & Gelman, 1998) or a specific manifestation of a particular Anglo-European individualistic interpersonal stance (e.g., Lillard, 1998).

METHOD

In a meta-analysis, conditions within studies, rather than individual participants or entire studies, comprise the unit of analysis (Glass, McGraw, & Smith, 1981). In our analyses, the proportion of children in a condition who demonstrated the target behaviors—correct false-belief judgments versus errors—was used as the essential dependent variable. (More precisely, we use the proportion of false-belief questions answered correctly, because many studies asked each child two or more false-belief questions.) A meta-analysis of such data is especially straightforward; many of the statistical problems that arise for other types of meta-analyses, which require the transformation and collapsing of various derived inferential statistics (e.g., *F* tests, *t* tests) (see Glass et al., 1981), can be avoided given this type of comparable descriptive data across studies.

In addition, we have been able to rule out aberrations due to sampling by supplementing our ordinary least squares estimates with bootstrap estimates of parameters and their standard errors. The details of these analytic methods are presented below in the Results section. Here, we present information on the studies and conditions included in the analyses.

Studies

The total potentially relevant literature is large. Therefore, the first task was to search pertinent journals, review articles, databases, and chapters for studies on several related topics, including false belief, mental representations, theory of mind, understanding mental states, belief-desire reasoning, and folk or naive psychology. The references of each citation were searched for additional articles. The studies generated in this fashion were supplemented by a search through several conference abstracts and by soliciting information from a variety of known researchers in this field (e.g., P. L. Harris, A. Gopnik, S. Baron-Cohen, J. W. Astington, J. H. Flavell, M. Siegal, M. Chandler, K. Bartsch, J. D. Woolley, and C. Lewis). Our efforts to find relevant studies were concluded in January 1998; no studies published or sent after that time were included in the analyses. In retrospect, we have discovered that several studies published before

1998 were omitted; given the size of the literature and its publication in a great many different journals this was inevitable.

Many studies that were encountered initially were not considered further because they reported no false-belief tasks or conditions. Some studies or conditions were not included because they focused only on autistic or delayed samples; the focus in our study was on normally developing children. In total 38 articles were examined and excluded for these reasons.

Not all studies or conditions purporting to include false-belief tasks with normally developing children could be included in the analyses. As already noted, different studies encompassed a wide variety of different task situations and questions. Most of these were reasonably straightforward, at least when sorted in relevant, measurable ways (e.g., deception versus none, puppets versus live protagonists). Some, however, were so different as to be irrelevant (e.g., conditions in which the real state of affairs was completely unknown, and hence, whether the protagonist's belief was false or true was indeterminable).

Moreover, for our procedures we needed comparably reported data on the dependent measure (proportion of correct false-belief judgments) and we needed the data to be reported separately for children of various ages (e.g., 3- versus 4-year-olds) and for separate tasks (e.g., locations versus contents tasks). Several studies collapsed their reported findings together across age or task features. For a number of these, more detailed data was sought and received from the authors, but this was not always possible or clarifications were not always forthcoming.

In total, our analyses encompassed 77 reports or articles including 178 separate studies and 591 conditions. A list of the number of studies and conditions appears in Table 1. These same 77 reports yielded 58 conditions that were excluded for the reasons described earlier (less than 10% of the conditions that were included).

One other exclusion was made. False-belief tasks typically ask not only the false-belief question (e.g., Where does Maxi think his chocolate is?) but also one or more control questions (e.g., Where is the chocolate really? Where did Maxi put his chocolate?). Performance on these control questions varies, but typically even the youngest children do reasonably well, and often only children who answer the controls correctly are then included in the final data. Conditions in which fewer than 60% of the children answered the control questions correctly or in which 40% or more of the initial subjects were dropped were excluded from our analyses. A total of nine conditions were excluded for these reasons; we assumed that these con-

ditions involved atypically difficult tasks or extremely confused children.

In total, a large majority of false-belief research was examined, representative of both published and unpublished false-belief studies in the field as of January 1998. The analyses of the data provided by these studies proceeded in several stages. In each, different subsets of the total data were used, as clarified in the Results section.

Coding

Each condition included in the analyses was coded for the dependent variable (proportion of correct responses to the false-belief question) and for a variety of features constituting the following independent variables:

1. Year of publication.
2. Mean age and number of participants in a condition.
3. Percentage of participants passing control questions, and percentage dropped from the research.
4. Country of participants: for example, the United States, United Kingdom, Austria, Japan, and so forth.
5. Type of task: Three levels of task type, which distinguished locations versus contents versus identity tasks (as described earlier).
6. Nature of the protagonist: Five levels that described the protagonist as a puppet or doll; a pictured character; a videotaped person; or a real person present in, or absent from, the current situation. Protagonists who were real, present persons were then also coded as either the self or another person.
7. Nature of the target object: Four levels that described the target object as a real object (e.g., chocolate), a toy, a pictured object, or a videotaped object.
8. Real presence of the target object: Two levels denoting whether, at the time the false-belief question was asked, the target object was real and present (e.g., chocolate was in the drawer, the crayon box contained candies) or not (e.g., Maxi's chocolate was used up and thus now absent, the crayon box was empty and thus contained no real object).
9. Motive for the transformation: Two levels capturing whether the key transformation (e.g., the change of location or the substitution of unexpected contents) was done to explicitly trick the protagonist (deception), or for some other reason including for no explicit reason at all.

Table 1 Listing of the Studies and Conditions Included in the Meta-Analysis

Authors	Year	Studies Reported	Studies Used in Meta-Analysis	Primary Conditions Included in Meta-Analysis	Nonprimary Conditions Included in Meta-Analysis	Total Conditions Included in Meta-Analysis
Astington, Gopnik, and O'Neill	1989	2	2	8	0	8
Avis and Harris	1991	1	1	0	6	6
Baron-Cohen	1991	1	1	0	1	1
Baron-Cohen, Leslie, and Frith	1985	1	1	0	1	1
Bartsch	1996	2	2	3	3	6
Bartsch, London, and Knowlton	1997	2	1	1	0	1
Bartsch and Wellman	1989	2	1	1	0	1
Berguno	1997	1	1	3	0	3
Carlson, Moses, and Hix	1998	3	2	3	0	3
Carpendale and Chandler	1996	2	2	13	0	13
Chandler and Hala	1994	4	4	14	0	14
Chen and Lin	1994	1	1	2	2	4
Clements and Perner	1994	1	1	4	0	4
Custer	1996	1	1	0	2	2
Dalke	1995	2	2	12	2	14
Davis	1997	2	1	2	1	3
Flavell, Flavell, Green, and Moses	1990	4	2	2	0	2
Flavell, Mumme, Green, and Flavell	1992	4	2	2	2	4
Freeman and Lacohee	1995	6	6	31	6	37
Freeman, Lewis, and Doherty	1991	5	4	15	7	22
Fritz	1992	2	2	10	2	12
Frye, Zelazo, and Palfai	1995	3	2	12	0	12
Ghim	1997	5	4	7	1	8
Gopnik and Astington	1988	2	2	30	12	42
Hala and Chandler	1996	1	1	8	0	8
Hala, Chandler, and Fritz	1991	3	1	3	0	3
Happe	1995	1	1	0	2	2
Harris, Johnson, Hutton, Andrews, and Cooke	1989	3	2	0	5	5
Hickling, Wellman, and Gottfried	1997	2	2	4	0	4
Hogrefe, Wimmer, and Perner	1986	6	4	16	0	16
Johnson and Maratsos	1977	1	1	4	0	4
Kikuno	1997	1	1	2	0	2
Koyasu	1996	1	1	9	0	9
Lalonde and Chandler	1995	1	1	6	0	6
Leekam and Perner	1991	1	1	2	0	2
Leslie and Thaiss	1992	2	2	4	0	4
Lewis, Freeman, Hagestadt, and Douglas	1994	5	4	0	6	6
Lewis and Osborne	1990	1	1	18	0	18
Lillard and Flavell	1992	2	2	3	0	3

(Continued)

10. Participation in the transformation: Three levels describing whether the child initially helped to set up the task props, engaged in actively making the key transformation, or only passively observed the events.
11. Salience of the protagonist's mental state: Four levels describing whether the mental state had to be inferred from the character's simple absence during the key transformation, whether the character's absence was emphasized and explicitly noted, whether the false-belief experience was demonstrated initially on the chil-

- dren themselves (e.g., the child initially discovered that the crayon box contained candies), or whether the character's mental state was explicitly stated (e.g., the child was told "Maxi thinks it is in the cupboard") or pictured in some fashion (e.g., via a thought bubble).
12. Type of question: Four levels denoting whether the false-belief question was phrased in terms of where the protagonist would look (or some other belief-dependent action the character might take), what he'd think or believe, what he'd say, or what he'd know.

Table 1 Continued

Authors	Year	Studies Reported	Studies Used in Meta-Analysis	Primary Conditions Included in Meta-Analysis	Nonprimary Conditions Included in Meta-Analysis	Total Conditions Included in Meta-Analysis
Mayes, Klin, and Cohen	1994	1	1	6	0	6
Mazzoni	1995	1	1	1	1	2
Mitchell and Lacohee	1991	3	3	7	0	7
Moore, Pure, and Furrow	1990	2	1	2	0	2
Moses	1993	2	2	0	10	10
Moses and Flavell	1990	2	2	4	0	4
Naito, Komatsu, and Fuke	1994	1	1	4	2	6
Perner, Leekam, and Wimmer	1987	2	2	9	5	14
Perner and Wimmer	1987	2	2	8	0	8
Phillips	1994	7	1	2	0	2
Riggs, Peterson, Robinson, and Mitchell	1998	4	2	2	0	2
Robinson and Mitchell	1992	5	2	3	0	3
Robinson and Mitchell	1994	3	3	5	0	5
Robinson and Mitchell	1995	6	4	6	2	8
Robinson, Riggs, and Peterson	1997	4	2	2	0	2
Robinson, Riggs, and Samuel	1996	3	3	8	0	8
Roth and Leslie	1998	2	2	4	6	10
Ruffman, Olson, Ash, and Keenan	1993	3	3	2	5	7
Russell and Jarrold	1994	3	3	4	4	8
Saltmarsh, Mitchell, and Robinson	1995	5	5	12	0	12
Seier	1993	1	1	0	1	1
Sheffield, Sosa, and Hudson	1993	1	1	4	0	4
Siegal and Beattie	1991	2	2	8	0	8
Slaughter and Gopnik	1996	2	2	3	1	4
Sullivan and Winner	1991	1	1	18	0	18
Sullivan and Winner	1993	1	1	12	0	12
Taylor and Carlson	1997	1	1	4	0	4
Vinden	1996	1	1	4	8	12
Wellman and Bartsch	1988	3	1	3	0	3
Wellman, Hollander, and Schult	1996	4	1	3	0	3
Wimmer and Hartl	1991	3	3	6	5	11
Wimmer and Perner	1983	4	3	16	0	16
Wimmer and Weichbold	1994	1	1	8	0	8
Winner and Sullivan	1993	1	1	6	0	6
Woolley	1995	2	2	7	1	8
Yoon and Yoon	1993	2	1	16	0	16
Zaitchik	1990	5	1	4	0	4
Zaitchik	1991	1	1	12	0	12
Totals		178	143	479	112	591

13. Temporal marker: Two levels capturing whether the false-belief question explicitly mentioned the time frame involved ("When Maxi comes back, which place will he look in first?") or not.

Because the way in which an individual condition was to be coded on the above measures was not always clear-cut, two independent raters coded 94 conditions representing 20 studies. Agreement (agreements divided by agreements plus disagreements) ranged from 92 to 100%. Over all codings, reliability averaged 97%. On some variables (year of publication, country, and motive for the transfor-

mation) agreement was perfect. The two lowest reliabilities were 92% (for percent passing control items and for salience of the protagonist's mental state). All disagreements were resolved through discussion.

RESULTS

Primary Conditions

As noted, the entire database for the meta-analysis contained information on 591 false-belief conditions. In the first wave of analyses, however, we included

only what we termed primary conditions. These were conditions in which (1) subjects were within 14 months of each other in age, (2) less than 20% of the initially tested subjects were dropped from the reported data analyses (due to inattention, experimental error, or failing control tasks), and (3) more than 80% of the subjects passed memory and/or reality control questions (e.g., "Where did Maxi put the chocolate?" or "Where is the chocolate now?"). Our reasoning was that age trends are best interpretable if each condition's mean age represents a relatively narrow band of ages; interpretation of answers to the target false-belief question is unclear if a child cannot remember key information, does not know where the object really is, or cannot demonstrate the verbal facility needed to answer parallel control questions. In most of the studies, few subjects were dropped, very high proportions passed the control questions, and ages spanned a year or less, so primary conditions included 479 (81%) of the total 591 conditions available. The primary conditions are enumerated in Table 1; they were compiled from 68 articles that contained 128 separately reported studies. Of the 479 primary conditions, 362 asked the child to judge someone else's false belief; we began our analyses by concentrating on these conditions. On average in the primary conditions, 3% of children were dropped from a condition, children were 98% correct on control questions, and ages ranged 10 months around their mean values.

In an initial analysis only age was considered as a factor. As shown in Figure 2, false-belief performance dramatically improves with age. Figure 2A shows each primary condition and the curve that best fits the data. The curve plotted represents the probability of being correct at any age. At 30 months, the youngest age at which data were obtained, children are more than 80% incorrect. At 44 months, children are 50% correct, and after that, children become increasingly correct. Figure 2B shows the same data, but in this case the dependent variable, proportion correct, is transformed via a logit transformation. The formula for the logit is:

$$\text{logit} = \ln\left(\frac{p}{1-p}\right),$$

where "ln" is the natural logarithm, and "p" is the proportion correct. With this transformation, 0 represents random responding, or even odds of predicting the correct answer versus the incorrect answer. (When the odds are even, or 1, the log of 1 is 0, so the logit is 0.) Use of this transformation has three major benefits. First, as is evident in Figure 2B, the curvilinear relation between age and proportion correct is

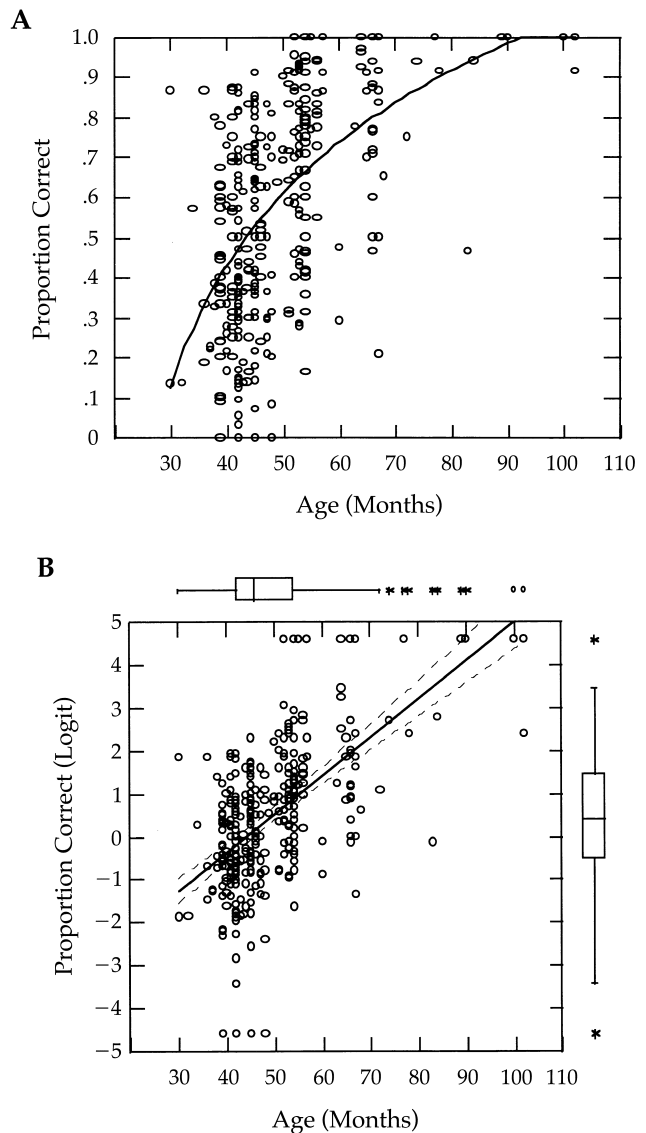


Figure 2 Scatterplot of conditions with increasing age showing best-fit line. (A) raw scatterplot with log fit; (B) proportion correct versus age with linear fit. In (A), each condition is represented by its mean proportion correct. In (B), those scores are transformed as indicated in the text.

straightened, yielding a linear relation that allows systematic examination of the data via linear regression; second, the restricted range inherent to proportion data is eliminated, for logits can range from negative infinity to positive infinity; and third, the transformation yields a dependent variable and a measure of effect size that is easily interpretable in terms of odds and odds ratios (see, e.g., Hosmer & Lemeshow, 1989).

The top line of Table 2 summarizes the initial analysis of age alone in relation to correct performance

Table 2 Summary of Meta-Analytic Results for the Primary Conditions

Variables	Main Effect	Interaction with Age	Effect Size ^a
Age: $F(1, 360) = 229.15, p < .001$			2.94 for 1 year
Nonsignificant			
Year of publication	$F(1, 325) = 2.65, p > .10$	$F(1, 324) = .96, p > .32$	
Type of task	$F(1, 359) = .63, p > .42$	$F(1, 358) = .64, p > .42$	
Type of question	$F(3, 357) = 1.96, p > .11$	$F(3, 354) = .81, p > .48$	
Nature of the protagonist	$F(4, 356) = 1.66, p > .44$	$F(4, 352) = 1.15, p > .33$	
Nature of the target object	$F(3, 357) = 2.49, p > .06$	$F(3, 354) = 2.35, p > .07$	
Self versus other	$F(1, 230) = 1.77, p > .18$	$F(1, 229) = 3.10, p > .07$	
Main effects			
Motive	$F(1, 359) = 14.27, p < .001$	$F(1, 358) = .30, p > .58$	1.90
Participation	$F(2, 358) = 5.91, p < .003$	$F(2, 356) = 1.25, p > .28$	1.96
Real presence	$F(1, 359) = 16.05, p < .001$	$F(1, 358) = .63, p > .42$	2.17
Salience	$F(3, 357) = 4.28, p < .006$	$F(3, 354) = .98, p > .40$	1.92
Country	$F(6, 345) = 10.42, p < .001$	$F(6, 359) = 1.04, p > .40$	Australia versus United States = 2.27 United States versus Japan = 1.48
Interaction			
Temporal marker	$F(1, 358) = 5.45, p < .02$	$F(1, 358) = 7.57, p < .006$	

^a Effect sizes are presented only for significant variables. Effect sizes were computed in odds ratios and represent the increased odds of being correct given a facilitating value of a variable, as explained in the text.

(transformed via the logit transformation). The far right column presents the measure of effect size: the odds of being correct increase 2.94 times for every year that age increases. At about 44 months of age, children are about 50% correct, or at 0 on the transformed measure; hence, at that age the odds of being correct or incorrect are even or 1.0 (Figure 2B). The effect size measure in Table 2 indicates the increase in the odds of being correct for a more advantageous value of a variable. A 2.94 increase for age means that for children who are 1 year older (i.e., 56 rather than 44 months of age) percent correct would be 74.6%. In terms of months, the effect size for age is 1.09 per month, or an increase from 50% to 52% correct from 44 to 45 months of age.

A meta-analysis must be guided by an organized set of questions and hypotheses. Our analyses were originally directed toward evaluating several baseline models of children's performance in false-belief tasks. The simplest model is a null model predicting random performance on false-belief tasks. Random performance at some age might suggest that children are confused, have few relevant systematic conceptions, or that the tasks tap little of children's understandings. Predictions of random performance contrast clearly with predictions of significantly below-chance performance (systematic false-belief errors) and significantly above-chance performance (systematic correct understanding of false beliefs). Developmentally, a baseline model predicts no developmental

change—for example, all ages evidence essentially random performance, or all evidence similar above-chance performance.

A central question for evaluating these baseline models is whether and where children's performance significantly differs from chance. Ninety-eight percent of all false-belief conditions in the studies sampled used two locations (Maxi's chocolate might be in the drawer or the cupboard), or two limited identities (the crayon box usually contains crayons but now contains candies), thus chance responding represents 50% correct (or a score of 0 in the transformed data). We calculated a 95% confidence band around the best-fit line shown in Figure 2B and then determined where performances within this band fell below or above 0. These confidence bands are also depicted in Figure 2B.¹ At younger ages—essentially 41 months (3 years, 5 months) and younger—children performed below chance, making the classic false-belief error. At older ages—essentially 48 months (4 years) and older—they performed above chance, significantly correct. As shown in Figure 2, then, average per-

¹ The confidence bands, here and throughout the analyses, are natural extensions of confidence intervals for the population mean, with the exception that confidence limits are formed for the conditional mean of the dependent variable at different combinations of the independent variables. These are simultaneous confidence bands that contain the error rate at .05, no matter how many combinations of independent variables are investigated (Draper & Smith, 1981).

formance changes rapidly during the period 3 to 4½ years from significantly wrong to significantly correct.

This age trend constitutes the foundation for analyses of the influence of various factors on performance in which we examined the effects of the 12 remaining independent variables (e.g., year of publication, type of task, nature of the target object, and so forth, as overviewed in the Method section). Again, consider in advance various patterns of results that might emerge. As depicted in Figure 3A, one pattern of results is that some additional variable (e.g., type of task—a comparison between locations versus contents tasks) would have no effect on the foundational age trajectory. Alternatively, as shown in Figure 3B, that variable may be significant, but only as a main effect that does not interact with age. That is, differences on that variable affect performance similarly at all ages. Finally, as depicted in Figure 3C, a focal variable could interact significantly with age. A variety of patterns might yield significant interactions with age, but the one shown in Figure 3C is especially relevant for discriminating early competence accounts from claims of conceptual change. To confirm early competence claims, task modifications should influence younger children's performance (and thus account for their poor performance relative to older children), and, in particular, some task version(s) should increase young children's responding to above-chance levels.

We tested for such patterns as follows. First, linear regression was used to screen for main effects and all two-way interactions with age. If interactions with age were absent, the interaction term was dropped from the regression and main effects were then tested. For those variables shown to have a significant effect, confidence bands were constructed around the obtained regression lines to determine where these lines differed from chance. Finally, these primary analyses were corroborated both by using more assumption-free bootstrapping methods and by examining the entire omnibus dataset of 591 conditions. Due to the large database, a conservative .01 level for significance was generally adopted. Because interaction effects are of particular theoretical importance, however, any interactions at the .05 level were considered significant as well.

As shown in Table 2, six variables did not significantly affect performance, neither by themselves nor in interaction with age; five were significant but failed to interact with age; and one significantly interacted with age.

Nonsignificant variables. Given 362 conditions, a failure to find significance with such a powerful test is noteworthy. For example, although year of publication

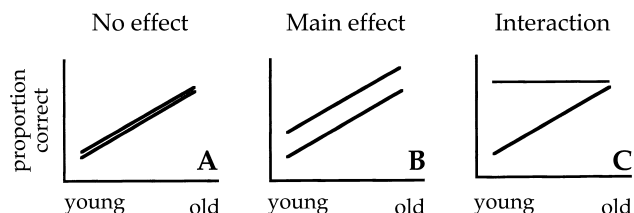


Figure 3 Three hypothetical patterns of results: (A) No effect, (B) main effect, and (C) interaction.

is of no theoretical interest, it provides a meta-analytic check for the possibility that an initial intriguing finding can shrink or disappear as later experiments are devised with better controls and procedures (Green & Hall, 1984). Some authors contend that early reports of young children's false-belief errors are plagued with just such problems. However, the meta-analysis shows that false-belief results from earlier studies are virtually identical to those from later studies.

More important, however, is that type of task, type of question, nature of the protagonist, and nature of the target object were all nonsignificant. To confirm graphically the absence of some of these variables as significant factors, Figure 4 shows the data for type of question and type of task. Type of question concerns whether the false-belief question is phrased in terms of what the character will think, know, or say, or where he will look. As can be seen in Figure 4A, these variations make little difference.

As noted, false-belief tasks come in three essential forms: change-of-location tasks, unexpected-contents tasks, and unexpected-identity tasks. In fact, unexpected-identity tasks are rare (used in only 44 of 362 conditions), and preliminary analysis showed that they did not differ from unexpected-contents tasks, which they closely resemble. Therefore, unexpected-identity tasks were collapsed with unexpected-contents tasks and Figure 4B thus plots the data for these two main task types, locations versus contents (including identity). As shown, these task types are remarkably equivalent.

In addition, the "medium" in which the false-belief task is presented has no significant effect. That is, it makes no difference if the protagonist is presented as a real person, a puppet, a doll, a pictured storybook character, or a videotaped person. Similarly, as long as a concrete target object is present at the time the false-belief question is asked, it makes no difference whether the object is a real item (a piece of edible chocolate), a toy (a small toy car), a picture of an object (a drawing of a piece of chocolate), and so forth. Most studies maintained concordances across protag-

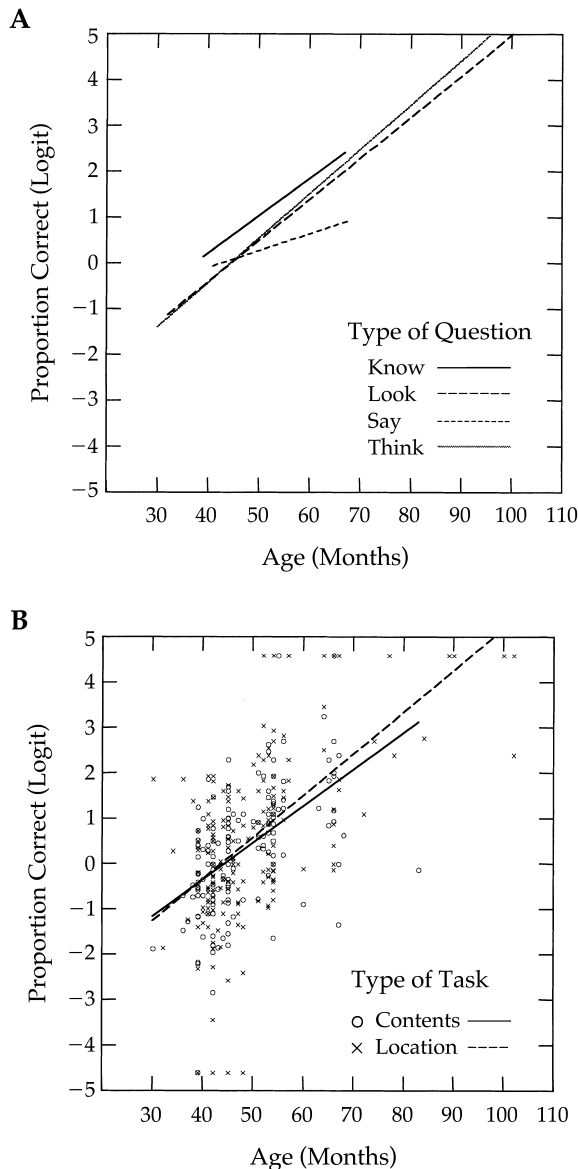


Figure 4 (A) Proportion correct versus Age \times Type of Question, and (B) proportion correct versus Age \times Type of Task.

onist and object. If a real person was used as the protagonist, then a concrete object or toy object was used as the object. Likewise, a doll or puppet protagonist was matched with a toy object and a drawing of a girl was matched with a drawing of a piece of chocolate. At least within these loose confines, children do not reason better or worse about real people versus dolls, or about real chocolate versus pictured chocolate.

Self versus other. The analysis for nature of the protagonist, described thus far, included only conditions in which the protagonist was someone else. Conditions in which the false-belief question was asked

about the child's own belief (self-conditions) were excluded for all the analyses summarized at the top of Table 2, with the exception of the line labeled self versus others. To compare whether false-belief judgments for self differ from those for other people, it was not appropriate to simply compare all self-conditions to all other conditions—because self-questions were asked for only a limited number of tasks. All self-questions are about a person—the child (rather than, e.g., a puppet or doll); all self-tasks are content or identity tasks, not location tasks. By their nature, all self-tasks require the child to experience an initial demonstration (e.g., to say a crayon box will contain crayons and then see that it contains candies instead). To insure comparability, therefore, the 117 self-conditions were compared with the 118 other-conditions in which the protagonist was a person, the task was a content or identity task, and the child viewed an initial demonstration of or participated actively in making the key transformation. As Figure 5 shows, in this analysis of 235 conditions, children's correct responses to false-belief questions for self versus other did not differ, and were virtually identical at the younger ages.

Figure 5 also shows that the essential age trajectory for tasks requiring judgments of someone else's false belief is paralleled by an identical age trajectory for children's judgment of their own false beliefs. Young children, for example, are just as incorrect at attributing a false belief to themselves as they are at attributing it to others.

Main effects. Five variables were significant as main effects, but did not interact with age. In essence, these factors can enhance children's performance, but in

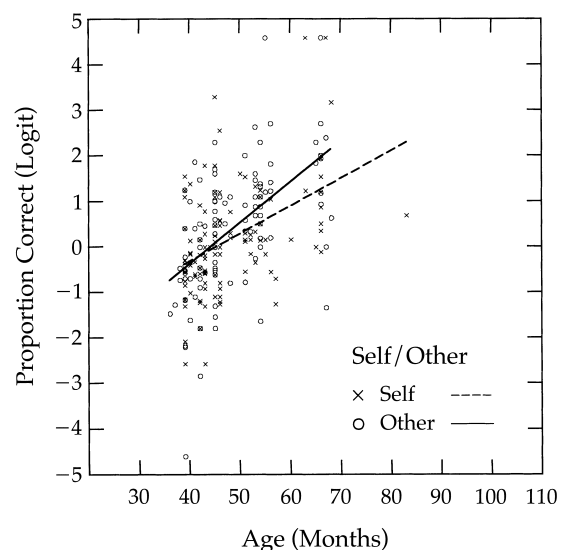


Figure 5 Proportion correct versus Age \times Self/Other.

doing so they leave the underlying developmental trajectory (from incorrect to correct performance) unchanged. Figure 6 graphically depicts the effects for motive for the transformation (“motive”), participation in the transformation (“participation”), salience of the protagonist’s mental state (“salience”), and real presence of the target object (“real presence”).

Motive concerned the motive presented to the child for the change of location or the unexpected contents. Sometimes a deceptive motive was explicitly stated (the chocolate was moved to trick the protagonist), sometimes it was not (either because no motive was stated—the crayon box just contained candies, or, rarely, a nondeceptive motive was mentioned—to put them away). As is clear in Figure 6A, a deceptive motive enhances children’s performance, and does so for children of all ages.

Table 2 presents *F* values for all the variables as well as effect sizes for significant variables. To reiterate, just as for age, this odds ratio measure indicates the increase in the odds of being correct for a more advantageous value of the variable: for example, for motive, the increase in odds of being correct for tasks with deceptive motives over those without deceptive motives. As shown in Table 2, this value for motive is 1.90, that is, a deceptive motive increases the odds of being correct by 1.90. In percentage terms, if children are 50% correct at 44 months without deception, then they are 66% correct with deception.

Participation by the child in transforming the target object is also important. Often children are essentially passive onlookers; for example, they watch as someone transfers Maxi’s chocolate from one place to another. However, in some tasks children help set up

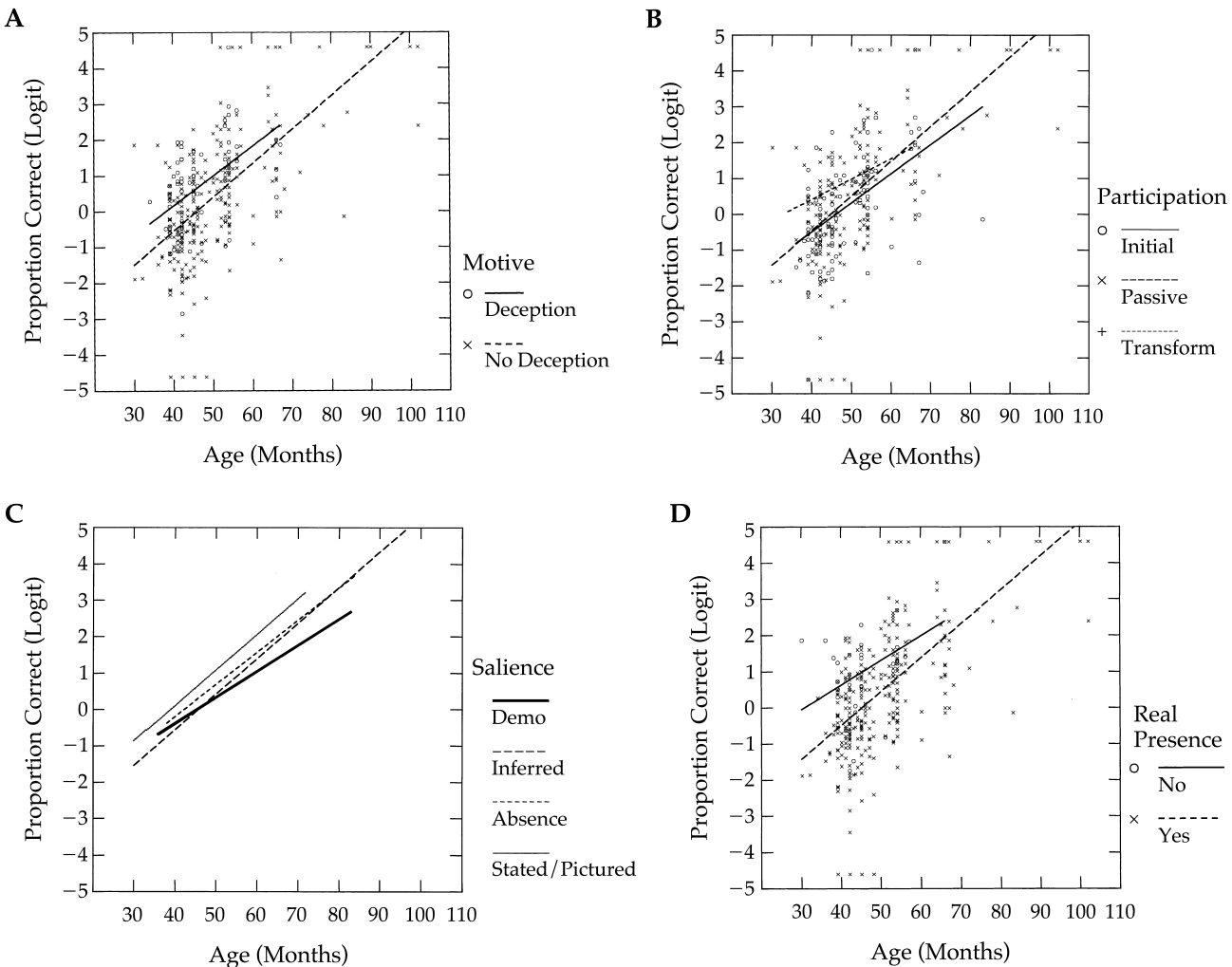


Figure 6 (A) Proportion correct versus Age × Motive; (B) proportion correct versus Age × Participation; (C) proportion correct versus Age × Salience; (D) proportion correct versus Age × Real Presence.

or manipulate the initial task situation or story props. Further, in some tasks the children themselves make the essential transformation, for example, moving Maxi's chocolate, or taking crayons out of a crayon box and putting in candies instead. As shown in Figure 6B, this last type of engagement—actively making or helping make the crucial transformation—positively influences children's performance at all the ages tested. As shown in Table 2, actively making the transformation increases the odds of being correct 1.96 over a baseline of being a passive onlooker. In percentage terms, if at 44 months children who are passive onlookers are 50% correct, then children who are actively involved in transforming the task materials are 66% correct.

The variables real presence and salience both arguably influence the extent to which the task focuses on mental-state information. Again, consider false-belief tasks as providing information about two realms of content: real-world contents and mental-state contents. One might attempt to make mental state more focal either indirectly by diminishing the salience of the contrasting real-world contents, or directly by enhancing the salience of mental states. The variable real presence describes the status of the target object at the time the false-belief question is asked—whether the true state of affairs is instantiated by a real and present object (e.g., Maxi's chocolate in the cupboard, or the candies in the crayon box) or not (e.g., Maxi's chocolate was removed from the drawer and eaten, and, thus, is not now real and present). As shown in Figure 6D, if the object is not real and present, children are more likely to answer correctly. With an effect size of 2.17, if children at 44 months are 50% correct with real and present objects, then they are 68% correct if the object is not real and present. Because this variable does not interact with age, however, both values of real presence leave the basic age trajectory unchanged. Moreover, even when there is no real and present object, the younger children do not perform at above-chance levels; 95% confidence bands around that line substantially overlap with zero. Thus, the youngest children may move from below-chance performance to at-chance performance, but only older children move to above-chance performance.

Comparably, salience of the protagonist's mental state is also significant. In this case, as shown in Figure 6C, most task variations are equivalent. For example, it does not matter if the protagonist is merely absent when the transformation is made and the mental state must be inferred, whether the character's absence is emphasized, or if the false-belief situation and experience is initially demonstrated on the children themselves (e.g., they discover the

crayon box contains candies). Yet, if the protagonist's belief itself is clearly stated or pictured, this significantly raises performance—for example, younger children move from below-chance to at-chance performance. Even if mental states are stated or pictured, however, young children do not achieve above-chance performance and the basic age trajectory remains the same.

The 44 conditions included as stated or pictured are worth "unpacking" further. In one type of stated or pictured condition, the protagonist's belief is stated (e.g., "Maxi thinks his chocolate is in the drawer.") and then the false-belief question asks where he will look (e.g., "Where will Maxi look for his chocolate?"). To be correct the child must at least recognize the implication of Maxi's thought for his behavior in a situation in which behaving according to belief means Maxi will search for his chocolate where it isn't. In a second type of stated or pictured condition, the protagonist's belief is again stated (e.g., "Maxi thinks his chocolate is in the drawer.") and the false-belief question asks what Maxi thinks (e.g., "Does Maxi think his chocolate is in the drawer or the cupboard?"). Arguably, in this type of task children could respond correctly by simply repeating back the earlier belief statement ("Maxi thinks his chocolate is in the drawer."). Finally, a third type of stated or pictured condition shows the protagonist's belief in pictorial form. For example, at the time Maxi puts his chocolate in the drawer the child is asked to choose a picture showing where the chocolate is at this point (and the picture is either kept in view or removed from view). Or, Maxi may be shown with a thought bubble depicting his belief (e.g., a thought bubble of chocolate in the drawer and not the cupboard). Intriguingly, all these variations produce largely the same results: they all are helpful, but none provide evidence for the type of interaction noted in Figure 3C.

Finally, the country of origin influences performance. Figure 7, which includes lines for the seven countries in which there were six or more total conditions, shows that at any one age, children from various countries can perform better or worse than each other. Nonetheless, children in all countries exhibit the same developmental trajectory. Conditions from the United States and the United Kingdom represent the largest sample (together contributing 48% of all conditions). Thus, using those countries as a baseline, children in Korea perform similarly; children in Australia and Canada perform somewhat better; and those in Austria and Japan perform somewhat worse. The effect size values shown in Table 2 present two extremes. If at

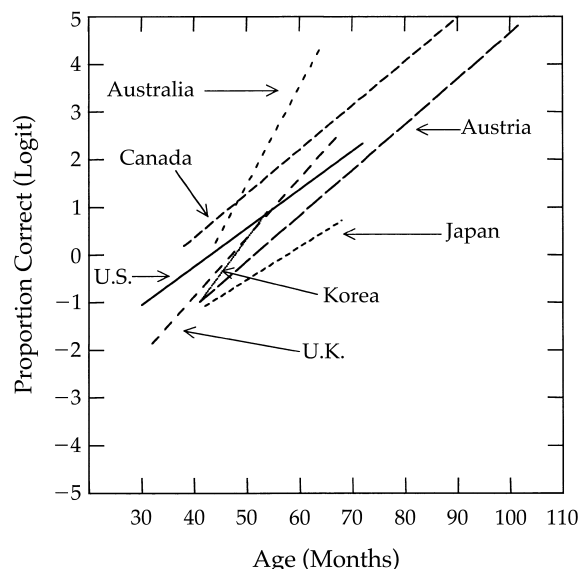


Figure 7 Proportion correct versus Age \times Country for the primary conditions.

44 months of age children in the United States are 50% correct, then children in Australia are 69% correct and children in Japan are 40% correct.

Interaction. One variable, temporal marking, interacts with age. This variable refers to whether the false-belief question emphasizes the time frame involved (e.g., "When Maxi comes back, where will he look first for his chocolate?"). As shown in Figure 8, only at older ages does including temporal information in the target question significantly increase correct performance. Temporal markers increase the length and complexity of false-belief questions. Conceivably this complexity may hinder, or at least fail to enhance, young children's performance, although older children seem to benefit from the clarity this additional information provides.

Bootstrap Analyses

Data for meta-analyses often fail to meet certain standard statistical assumptions. For example, in our analyses some studies contributed 10 or more conditions to the dataset but others contributed only one or two conditions, making independence assumptions problematic. Bootstrap analyses are more assumption free.

The basic idea of the bootstrap is to replace the theoretical normal distribution with the empirical distribution formed by the data themselves, N observations (Freedman & Peters, 1984). One then repeatedly samples with replacement (B samples of size N) from this empirical distribution, each time computing the

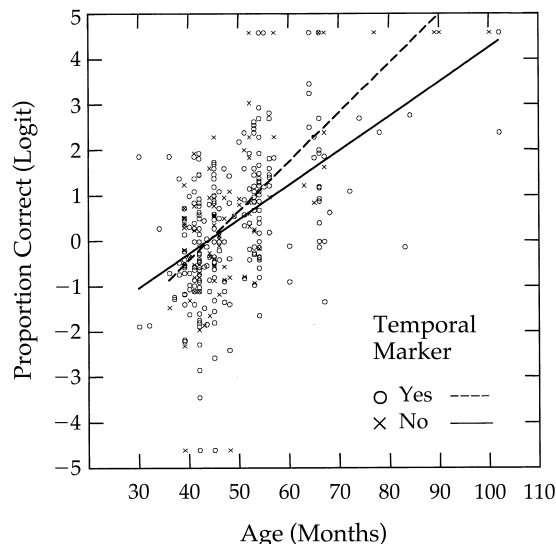


Figure 8 Proportion correct versus Age \times Temporal Marker.

coefficients of the regression model and storing the results. When the procedure is completed, there are B estimates of each parameter, each computed on a randomly selected set of N observations. The mean of these B estimates is the bootstrap estimate of the parameter; the standard deviation of the B estimates is used to estimate the standard error of each parameter.

The key analyses in this study were confirmed by conducting parallel bootstrap analyses (where $N = 362$ and $B = 1,000$). For example, consider the regression model underlying Figure 2B. If only age is used to account for false-belief judgments, then $R^2 = .39$, and the following coefficients are obtained (with standard errors in parentheses)

$$\begin{aligned} \text{proportion correct (transformed)} = \\ -3.96 [\text{constant}] + .09 [\text{age}]. \\ (.305) \quad (.006) \end{aligned}$$

Using bootstrap procedures to calculate the parameters of this model 1000 times yielded mean values of -3.98 for the constant and $.09$ for age, with the standard deviation for these parameters being $.303$ and $.006$, respectively. In short, this bootstrap analysis closely replicates the earlier finding in all respects.

Bootstrap analyses confirmed all the effects reported above. In particular, bootstrap analyses confirmed the absence of interactions between age and any other significant factor, except temporal marking.

Omnibus Analyses

The findings from primary conditions were also corroborated by conducting similar analyses on the

entire omnibus dataset. For example, the analysis of age, depicted in Figure 2, includes the data from 362 primary conditions; the parallel omnibus analysis included 453 conditions. Age was significant in the omnibus analysis, replicating the primary analysis. The regression equations are closely similar:

$$\text{DV (primary)} = -3.96 [\text{constant}] + .090 [\text{age}]$$

$$R^2 = .39;$$

$$\text{DV (omnibus)} = -3.51 [\text{constant}] + .090 [\text{age}]$$

$$R^2 = .28.$$

The principal difference is that the omnibus analysis is noisier, yielding a lower R^2 and consequently accounting for less of the overall variance.

Similarly, each variable that was significant in the primary analyses was re-analyzed, using the omnibus dataset. Results closely replicated the primary analyses, although in each case they proved noisier and thus accounted for less overall variance. Specifically, results for motive, real presence, participation, salience, and country again evidenced main effects and no interactions. Temporal marking evidenced the same interaction that was apparent in the primary analyses (shown in Figure 8). We prefer the primary analyses over the omnibus ones because, methodologically, the primary analyses focused on studies with better controls and more complete data, and, analytically, the primary analyses provided a better accounting of the variance. Nonetheless, the close similarity of the results from the primary and omnibus analyses further attests to the robustness of the findings.

The omnibus data allow for the most comprehensive look at the influence of children's country of origin on their false-belief reasoning. It is reasonable that during attempts to tailor tasks to different cultural materials, nonstandard tasks might result. Figure 9, in conjunction with Figure 7, therefore, presents more complete information about country. The line for the United States in Figure 9 closely replicates the U.S. line in Figure 7 for the primary analysis, and thus serves as a baseline for other comparisons. The other lines in Figure 7 depict performance by children in Western and non-Western literate, developed countries. The omnibus data, however, include children from two nonliterate, more traditional communities: the line in Figure 9 for Africa depicts responses from hunter-gatherer Baka children (Avis & Harris, 1991) and the line for Peru depicts data from Quechua-speaking Peruvian Indians (Vinden, 1996). Just as in the primary analyses, country is clearly significant but does not interact with age. For these two cultural communities, just as for the others shown in Figure 7, children's false belief performance increases across

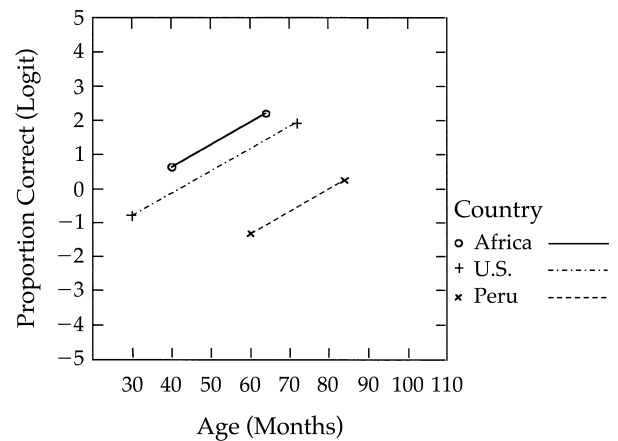


Figure 9 Proportion correct versus Age \times Country for three different countries within the omnibus conditions.

years in equivalent age trajectories, although at any one age children from different countries and cultures can perform differently.

Multivariate Accounting

To summarize, four task variables help the performance of younger (as well as older) children—motive, participation, salience, and real presence. One non-task variable, country, also included values that enhanced performance. In all the analyses reported thus far, younger children (30–40 months of age) remain below or at-chance level. It remains to assess, however, if a combination of task enhancements might enable very young children to perform systematically above chance level. In addition, it is important to determine what combined model best predicts children's false-belief performances. Finally, testing multiple factors in a combined model is important because the significant influence of one factor may disappear if the other factors are controlled for as well. For example, the effects for country may conceivably be due to investigators in one country exclusively using more helpful task variations; or the apparent benefits of active participation may disappear if motive is also included in the analysis (if, for example, almost all active participation tasks also included deception).

We tested combined models on primary conditions because regression predictions were more precise for primary conditions rather than the omnibus set (which, to reiterate, consistently yielded similar, but statistically noisier, results). The first model tested included all variables that significantly enhanced young children's performance in the earlier two-way

analyses—country, motive, participation, salience, real presence, and age. With these factors included, the multiple $R = .74$, and $R^2 = .55$. Hence these factors accounted for 55% of the variance in children's correct false-belief judgments. Again, bootstrap analyses closely replicated this regression model.

Because country is a subject variable not under experimental control, it was excluded in the next model tested. In this model all factors except participation significantly and independently contributed to the overall prediction, multiple $R = .68$, $R^2 = .47$. The final model therefore dropped participation and included motive, salience, real presence, and age, multiple $R = .68$; $R^2 = .46$.²

Using this final model three predicted sets of effects were then examined, as shown in Figure 10. The "best-effects" alternative shows predicted performance when values of all variables are maximally enhanced. Therefore, for motive, the best-effects model was based on framing the transformation in terms of explicit deception of the protagonist; for salience, this model was based on explicitly stating or picturing the protagonist's (false) mental contents; and for real presence, this model was based on the absence of a real target object at the time of the child's judgment. The "worst-effects" alternative examined predicted performances in the presence of detrimental values of these variables. For example, for motive the worst-effects model was based on the absence rather than the presence of explicit deception. The "no-effects" predictions were based on neutral or average values of each variable. As shown in Figure 10A, the no-effects prediction closely mimics the overall age changes revealed in the initial analysis of age alone (shown in Figure 2B). Most important, the best-effects prediction in Figure 10A is identical in slope to the no-effects line; even the best-effects combination of these enhancing variables does not differentially enhance younger children's performance. Moreover, a 95% confidence band that is fit around the best-effects line substantially overlaps with 0 or chance performance. These confidence bands are shown in Figure 10B. Specifically, the best-effects combination does not yield significant above-chance performance at the youngest ages; only at 40 months and older does that

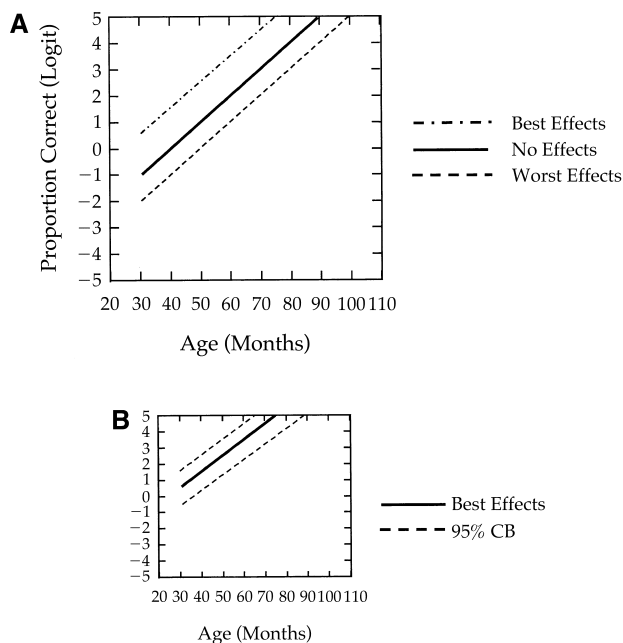


Figure 10 (A) Prediction intervals for final model. (B) Confidence bands for the best effects predictions.

combination of variables produce significant above-chance performance.

DISCUSSION

The integration of information across the many studies of children's understanding of false beliefs has become a necessity. Yet, the numerous studies make integration difficult, thereby encouraging reviewers to ignore some troublesome results or view the findings through the lens of prior theoretical commitments. A meta-analytic, quantitative integration of the research can overcome these difficulties.

Meta-analyses, of course, have potential weaknesses. For example, they can be undermined if publication biases work against certain, usually nonsignificant, findings. However, the literature on false beliefs is replete with studies of above-chance, chance, and below-chance findings because the occurrence of false-belief errors has become a contentious issue. Furthermore, in our analyses, we took care to include unpublished as well as published research.

More generally, meta-analyses are dependent on the quality of the accumulated studies to date. We have addressed concerns about quality by concentrating on a "primary" set of conditions where, on average, 98% of subjects passed relevant control tasks and few subjects were dropped. Moreover, a confirmatory analysis with a larger "omnibus" set of conditions

² This combined model arguably provides the most precise estimates of effect size for these key variables, because the influence of any one variable is controlled for the influence of other variables in this model. The effect size values generated from this model were: age, 3.32 (for each increase of 1 year); motive, 2.10; real presence, 2.63; and salience, 1.80. Comparing these values to those in Table 2 shows close agreement as well as some modest adjustments. Note that in this combined model the effect of age is not attenuated (as expected by an early competence hypothesis).

produced results almost identical to the original analysis. Reservations about meta-analytic findings also arise because of concerns with the appropriateness, assumptions, and robustness of various techniques for pooling indirect measures of effect significance (e.g., *p* values). The analyses in this study were more straightforward; we analyzed proportions of correct answers, rather than derived measures from statistical treatments that vary across studies. Finally, our analyses encompassed a great number of task variations. Yet when these variations were organized into a systematic set of factors, the results clustered consistently with the exception of only a few outliers. This is testimony to the robustness of the basic phenomena, the tasks used to assess it, and our meta-analytic procedures.

Using these meta-analytic procedures, a variety of substantive results emerged. The combined-effects models made it possible to predict approximately 50% of the variance in false-belief judgments. This is a strong accounting of the variance, because in meta-analyses there are a host of factors that typically attenuate findings; in this case, there were differences across studies in experimenter characteristics and expertise, differences in exact stimuli and procedures used, linguistic differences across a variety of countries, and differences in child-rearing cultures sampled. In what follows, the substantive results are considered in a series of ordered sets, and the implications of the results with regard to current proposals about theory-of-mind development are addressed.

Development

The basic finding in this study is the presence of a substantial effect for age in every analysis. Correct performance significantly increases with increasing age; in some cases, correct performance increases from chance to above chance, but in most cases it increases from below chance to above chance. Such findings clearly support the initial claims of substantial development during these preschool years, and contradict recent suspicions that developmental change is nonexistent (Mitchell, 1996) or is confined to only a few standard tasks that are unusually demanding (Chandler et al., 1989).

"Noneffects"

Many potentially relevant variables (e.g., nature of the protagonist, nature of the target object, type of question, and type of task) systematically failed to affect children's age-related performances. These results confirm those of several individual studies that

have failed to find differences for these variables. However, detecting a lack of differences when comparing small samples of 12 to 20 children is problematic. In contrast, because our meta-analysis encompassed 350 to 500 different conditions representing more than 4,000 children, it is certainly powerful enough to detect differences on these variables, if they existed.

The null findings of this study have important methodological implications. Knowing that valid and comparable assessments of false-belief performance are uninfluenced by a variety of task specifics, experimenters can confidently vary their tasks over an extended set of possibilities for ease of presentation and to achieve other experimental contrasts. More crucially, this lack of effects is of theoretical importance as well. That children's false-belief judgments are systematically unrelated to such task variations increases the likelihood that their judgments reflect robust, deep-seated conceptions of human action, rather than task-specific responses provoked by the special features of one set of materials or questions. Specifically, the irrelevance of these task and procedural variations increases the likelihood that children's performance is systematically dependent on the one thing that does not vary, namely their conception of belief states.

For adults, the task variations mentioned above are conceptually equivalent or irrelevant; it is a substantive finding that young children treat them as equivalent as well. Consider the notion of belief that is targeted by standard false-belief tasks. The beliefs involved are determined by informational access—whether or not Maxi saw the chocolate being moved—but should be uninfluenced by irrelevant individual differences in the target protagonists, such as their age, gender, and their real life presence versus video or storybook nature (Miller, *in press*). The meta-analytic findings show that even young children appropriately understand the irrelevance of several protagonist differences for an understanding of belief.

Early Competence versus Conceptual Change

As noted in the Introduction, our findings speak to an important divide between two general accounts of young children's poor false-belief performance. One account argues for conceptual change, that is, developmental changes in performance on false-belief tasks reflect genuine changes in children's conceptions of persons. A contrasting account argues for early competence, that is, even young children have the necessary conception; their poor task performances reflect instead information-processing

limits, unnecessarily demanding tasks, or confusing questions. Fodor (1992, p. 284), for example, argued that something like belief-desire psychology is innate: "The experimental data offer no reason to believe that the 3-year-old's theory of mind differs in any fundamental way from adult folk psychology. In particular, there is no reason to suppose that the 3-year-old's theory of mind suffers from an absent or defective notion of belief." Chandler et al. (1989, pp. 1263–1264) outline two competing opinions about the understanding of mind, and thus about the understanding of beliefs. There are the "boosters," who "advocate the . . . early onset view," and the "scoffers," who share a "delayed-onset perspective." Chandler and colleagues (p. 1266) champion the booster point of view and claim that the "standard assessment task . . . conflates the . . . capacity to entertain beliefs about beliefs with the altogether different ability to comment on this understanding, and is more tortuous and computationally complex than is necessary or appropriate."

Empirical confirmation of early competence accounts requires that there be some version of the target task that indeed demonstrates enhanced performance by young children, when task limitations have been eliminated or reduced. Our findings show that several task manipulations do increase young children's performance: framing the task in terms of explicit deception or trickery, involving the child in actively making the key transformations, and highlighting the salience of the protagonist's mental state or reducing the salience of the contrasting real-world state of affairs, all help young children to perform better.

Early competence accounts require more than just improved performance, however. First, such accounts require that relevant task manipulations differentially enhance young children's performance. This requirement derives from the essential claim that such task factors mask early competence, thereby artifactually producing apparent developmental differences on the target tasks. Second, such accounts require demonstrations of above-chance performance. A task manipulation that raises young 3-year-olds' performance from below-chance to chance performance may be methodologically important in that the manipulation reduces features that systematically lead children to choose an incorrect response option. Random at-chance performance is, however, neither systematically misled nor systematically informed; it provides no evidence of correct, conceptual understanding. Third, demonstrations of above-chance performance must extend beyond children at an intermediate, transitional age. If children at an intermediate

age are aided, but younger children still genuinely fail the task, then the timing of a developmental change, but not its absence versus presence, may be at issue. Thus, for early competence accounts, task variations should interact with performance in a specific developmental pattern.

None of these three pieces of evidence for an early competence view were sustained in the meta-analysis: older children as well as younger children were aided by certain task manipulations; no set of manipulations boosted younger children's performance to above chance; and only one variable, presence or absence of temporal marking in the test question, interacted with age, but not in an early competence pattern. Even a combined-effects model, statistically assembling the most powerful package of helpful task manipulations, failed to produce above-chance performance in the youngest children and failed to interactively change the shape of the basic developmental pattern of performance across these years. It is important to note that the finding of an interaction between age and temporal marking demonstrates that the meta-analysis was able to detect interactions in the data, if they were present.

Conceptual change accounts also require specific empirical findings. To begin with, a task is needed that plausibly assesses a target conceptual understanding, and performance on that task must change from incorrect to systematically above-chance judgments with age. Confidence in both good and poor performance on the task must additionally be bolstered by correct judgments on control tasks or on control questions that demonstrate memory for key information and grasp of the task format. False-belief tasks, especially those included in the primary analyses, fit these criteria. As shown in Figure 2, for example, performance on such tasks dramatically changes with age, and 98% of all children in primary conditions pass relevant control tasks.

Conceptual change accounts require more than simply age changes on a well-controlled standard task, however. They are also subject to a general multi-method approach to construct validity: a variety of tasks, all conceptually similar but varying in their task specifics, should lead to similar developmental changes. That is, the more the same developmental change is demonstrated in a variety of conceptually similar tasks (which vary in their specific features and demands), the less likely it is that the change is due to specific information-processing strategies tied to special task formats, or to limitations in understanding or using particular response formats. In this regard it bears repeating that in the meta-analysis a wide variety of task materials and formats yielded equivalent

developmental performances. The response requirements of the tasks included in the meta-analysis also varied widely: in some tasks the child could answer by simply pointing, or could answer a question about the character's behavior ("Where will he search?") or his mental state ("What will he think?"). Indeed, although nonverbal tasks were not available at the time we conducted the meta-analysis, similar age trajectories are obtained when the task is completely nonverbal (Call & Tomasello, 1999).

Conceptual change accounts do not require that task demands and information-processing limitations have no influence; on the contrary, conceptual understandings can only be manifest in specific tasks via processes of information utilization and expression. Conceptual change accounts do require that task demands and processing limitations not completely account for performance across age so that control of such factors does not eliminate a larger developmental pattern. The meta-analysis showed that a number of task manipulations do influence children's false-belief performances. Although these factors are important in their own right, they nonetheless leave the basic developmental pattern unchanged, even when considered in combination.

Meta-analytic results capture significant trends across studies. Of course, not every individual study accords with the group trends. Of particular relevance in the current case is whether there are any individual studies that reported clear early competence patterns of data. Most important is whether any studies reported above-chance performance from very young children (e.g., 2-year-olds). Such results would appear in the top left quadrant of Figure 2A. Inspection of Figure 2A shows that two data points stand out from the rest as indicating highly correct performance (better than 85% correct) from 30- and 36-month-olds. These data come from a study by Sheffield, Sosa, and Hudson (1993) using a "simplified" task in which there was no real presence of the target object and the character's mental state was explicitly stated.

Why might Sheffield and colleagues' conditions emerge as outliers (e.g., was the task that they used especially sensitive, or artifactually easy)? In Sheffield et al.'s simplified task, children see a red and a green box and discover that both are empty. Then Cookie Monster enters and explicitly says, "I think my cookie is in the red box." Immediately after this statement the child is asked the false-belief question: "Where does Cookie Monster think his cookie is?" To be correct on this task, therefore, the child need only repeat back what Cookie Monster just said, poten-

tially without any real understanding of belief at all. Arguably, therefore, Sheffield et al.'s findings represent artifacts of this particular task format.

In studies other than Sheffield et al. (1993), correct performance also could be achieved, potentially, by children merely repeating what was told them. In Flavell, Flavell, Green, and Moses (1990), the child knows that a cup hidden behind a screen is blue. Another person, Ellie, who enters the room, cannot see the cup, but explicitly states, "I think the cup is white." Then the child is asked, "Do you think the cup is white or blue?" and finally is asked, "Does Ellie think the cup is white or blue?" In the Flavell et al. study, 3-year-olds systematically erred by saying Ellie thinks the cup is blue. When children err in this fashion it is remarkable, because to err they avoid a very simple response strategy—just repeat Ellie's own words. Incorrect answers, therefore, may informatively indicate problems with conceptualizing false beliefs. Correct responses with such a procedure are much less interpretable, however, because, when correct, the child may indeed be simply parroting what they just heard.

A related issue concerns the results of the combined models. The best-effects model shows predicted responses if children are presented with tasks that combine several key enhancing features. No studies, however, have included a task with all four significantly enhancing factors: that is, a task in which the salience is pictured/stated, the motive is deceptive, the object is not real and present, and the child participates actively in the transformation. Sheffield et al. (1993) included only two of these four enhancing features, and only one condition in the meta-analysis included three of these factors in a task for young preschoolers. Woolley (1995, Experiment 1) tested young 3-year-olds ($M = 39$ months) with a task in which the motive was deceptive, the object was not real and present, and the child actively participated in the transformation. In Woolley's study, young children performed below chance: 14 correct responses out of 38 false-belief questions, or 37% correct. Thus, the predictions of the combined-effects model are both derived from and are largely consistent with observed results from individual studies—very young children typically fail to perform significantly above chance, even when tested on tasks that combine task manipulations that enhance performance.

With regard to a general contrast between conceptual change and early competence accounts, then, the meta-analysis suggests that an important conceptual change in children's understanding of persons is taking place between the ages of 2½ and 5 years. Against this general background, the detailed results of the

meta-analysis also speak to several more specific theoretical proposals.

Chandler's and Leslie's Accounts

Recall that Chandler (1988; Chandler et al., 1989) proposed that early competence at false-belief judgments is masked by unnecessarily demanding features of the standard tasks. In contrast to standard tasks, Chandler argued, tasks framed in terms of explicit deception, which enlist the child in concocting the deceptive circumstances and minimize verbal demands by having the child point or nonverbally arrange props, would better reveal young children's early competence. In their research, tasks employing these features have been found to enhance young children's performance (e.g., Chandler et al., 1989; Hala, Chandler, & Fritz, 1991), although in other studies this has not been the case (e.g., Sodian et al., 1991). The meta-analysis showed that these features—especially deception and active participation—indeed aid performance. The pattern of findings for these factors, however, failed to fit an early competence model more specifically. To reiterate, these task manipulations, although helpful, do not alter the shape of the general developmental trajectory and do not raise the youngest children's performance to systematically above-chance performance.

Leslie also proposed a specific early competence account (e.g., Leslie & Roth, 1993; Roth & Leslie, 1998) in which understanding persons' mental attitudes or states, such as the "belief that X is so," is the result of a special Theory-of-Mind Mechanism (ToMM) that is activated early in development. Performance in any task situation, however, depends not simply on ToMM but also on a Selection Processor (SP) that can limit the application of ToMM in any specific case. According to this account, standard false-belief tasks place large demands on SP and hence mask, rather than reveal, young children's Theory-of-Mind competence. In contrast, nonstandard tasks can reduce these demands thereby revealing early ToMM competence. Thus, Leslie (Leslie & Roth, 1993, p. 99) notes approvingly that "three-year-olds do succeed on some non-standard false-belief tasks" (see, e.g., Mitchell & Lacohee, 1991; Roth & Leslie, 1998; Wellman & Bartsch, 1988). The arguments above against early competence accounts more generally apply to Leslie's specific account; manipulations of the type that Leslie endorses do aid performance, but across numerous studies in the meta-analysis such manipulations did not systematically produce an early competence pattern of results.

Conversational Accounts

Several authors claim that young children have special difficulties with the verbal-conversational aspects of standard false-belief tasks. As conversational skills improve, children become able to reveal their pre-existing conceptual competence. For example, beginning with Lewis and Osborne (1990), investigators have worried that the typical false-belief question, ("Where will Maxi look for his chocolate?") is unclear. Perhaps young children interpret the question as, "Where will Maxi end up looking, after he first fails to find his chocolate?," thereby helpfully resolving any unclarity in favor of Maxi's success. Siegal (Siegal, 1997; Siegal & Beattie, 1991) more comprehensively considers several potential conversational problems. For example, standard tasks typically include repeated questioning of the child. In an unexpected contents task, for example, the child is asked, "What do you think is in here?"; then "What is really in here?"; followed by "What will George think is in here?" With repeated questioning, young children may alter their answers, producing erroneous responses albeit conceptually understanding false-belief problems. Switching answers for conversational reasons might be especially problematic when children are given multiple false-belief trials in a row.

The meta-analysis showed that revising the standard tasks to overcome potential conversational difficulties can influence children's performance, as shown for the variable "temporal marking." A temporally marked question such as "Where will Maxi first look for his chocolate?" clarifies that the question is about Maxi's first search and not a later successful search. Providing temporal clarification of this type does aid performance, but, opposite to what an early competence account would require, temporal marking fails to enhance very young children's performance, enhancing judgments only for older children.

What about repeated questioning? Children's consistency in responding to multiple false-belief trials is relevant here. In many studies, children receive a single false-belief trial. But just as often children receive two or more false-belief trials (i.e., they are presented with several equivalent false-belief tasks in a row in which only the characters and objects are changed across trials). The dependent variable in the present analyses (proportion correct) was summed across false-belief trials of any equivalent type. Summing across multiple trials was deemed appropriate for two reasons. Practically, research articles often do not break down proportion correct on separate trials, only providing information summed across trials.

Empirically, when studies do provide information about children's consistency across trials, they are typically found to be highly consistent. Within the 178 studies included in the present analyses, there were 52 reports of consistency across trials. This data took several forms but could be converted across studies into the proportion of times that responses on a second trial agreed with those on the first trial (proportion of agreement, or PA). Across these 52 reports, mean PA was .84 ($SD = .14$); on average, children gave identical responses to two or more similar false-belief trials 84% of the time. Consistency was correlated with age, $r(51) = .28, p < .05$, but even in conditions in which children's mean age was less than 44 months (19 of the 52 reports) the mean PA was .81.

These results speak to conversational concerns that standard tasks encourage answer switching and thus mask early competence. Children are quite consistent on multiple false-belief tasks of the same type. Recall also that children's performance does not differ if the question is asked in terms of "look," "think," or "say"; or if responses require simply pointing at one location or another, yes-no answers, or longer verbalizations. In short, the findings argue against straightforward early competence conversational accounts. Of course, the link between conversation and conception is not merely that the former may mask the latter. Conversational experiences may well contribute to children's developing conceptions of mind in some form or other (e.g., Astington & Jenkins, 1999). The meta-analysis cannot speak to this proposal, but several other findings support an important role for conversational experiences in stimulating children's understanding of false belief. Deaf preschool children of hearing parents are unlikely to engage in conversational exchanges about persons' mental states and also do poorly on false-belief tests (e.g., Peterson & Siegal, 1995). Dunn et al. (1991) found that 2-year-olds in families who frequently talked about emotions, particularly their causes, evidenced a more sophisticated understanding of belief and false belief at 4 years of age. Early family conversations about desire predict later understanding of belief and false belief as well (Bartsch & Wellman, 1995).

Children's understanding of beliefs can also be assessed via their everyday conversations about people, using such terms as think and know (Bartsch & Wellman, 1995; Moore, Furrow, Chiasson, & Patriquin, 1994; Shatz, Wellman, & Silber, 1983). Conversational contrastives, such as "I thought this was a crocodile, but it's an alligator," in which the child contrasts belief and reality specifically, convincingly reveal some understanding of false beliefs (Bartsch & Wellman,

1995). Conversational analyses consistently uncover false-belief contrastives of this type in children between the ages of 3 and 4. Comparing such conversational evidence with findings from standard false-belief tasks has led some to conclude that conversational understanding of false beliefs surpasses laboratory understanding, and even that conversational data reveal early competence rather than conceptual change. The findings of the meta-analysis help to resolve this apparent discrepancy: Laboratory tasks that include such helpful features as deception, participation, and absence of a real and present target object can significantly increase correct responding in 3-year-old children. Arguably, conversational and laboratory situations that are comparable on these types of features would yield similar data. More important, comprehensive longitudinal conversational data (Bartsch & Wellman, 1995) show a clear developmental trajectory from no understanding to understanding of belief and false belief, just as do the meta-analytic data.

Executive Function

Recently, investigators have increasingly focused on the relation between "executive functioning" and developments in understanding of mind and beliefs (e.g., Carlson, Moses, & Hix, 1998; Frye, Zelazo, & Palfai, 1995; Russell, 1996). Executive functioning encompasses several constructs including planning, response inhibition, and cognitive flexibility that may themselves be quite heterogeneous (Zelazo, Carter, Reznik, & Frye, 1997). Not surprisingly, then, there are several theoretical variations on how executive functioning and theory of mind may interrelate. An initial claim (Russell, Mauthner, Sharpe, & Tidswell, 1991), however, and a claim considered by all executive functioning accounts, constitutes an early competence proposal: that young children's difficulties on Theory-of-Mind tasks (especially false-belief tasks) stem from an inability to demonstrate conceptual knowledge due to executive functioning limits coupled with the demands of false-belief tasks. To illustrate, consider Carlson and Moses' proposals (Carlson et al., 1998; Carlson & Moses, *in press*).

Carlson and Moses focus on inhibitory control—the capacity to suppress actions or thoughts that are irrelevant to performance on some task. Inhibitory control develops markedly in the preschool years (Frye et al., 1995; Gerstadt, Hong, & Diamond, 1994; Kochanska, Murray, Jacques, Koenig, & Vandergeest, 1996); and inhibition of salient experiences and typical or prepotent responses seems to be necessary for correct judgments on false-belief tasks. Typically, one points to where an object is; to be correct on a false-

belief task can require pointing to where the object is not. Typically, one uses language to describe what is true; to be correct in false-belief tasks, one describes something false. More generally, as noted before, it is possible to describe false-belief tasks as encompassing two realms of content: real-world contents (that the chocolate really is in the cupboard) and the contents of the mind (that Maxi thinks the chocolate is in the drawer). Correct responses require suppressing or inhibiting reference to reality and referring instead to mental contents (and indeed to mistaken, false, or misguided mental contents).

As noted in the Results section, attempts to decrease the salience or prepotence of real-world contents have produced tasks in which the value on the variable real presence is not real and present. For example, Maxi's chocolate is removed from the drawer but then eaten or destroyed (rather than moved to the cupboard). In this case, because there is no real chocolate, responding correctly to the question, "Where will Maxi look for the chocolate?" does not require overcoming the prepotent response of referring to where the chocolate really is. In the meta-analysis, tasks in which a concrete real object is absent, and thus the salience of real-world contents is diminished, were shown to improve performance. Such task manipulations, however, do not alter the basic developmental trajectory of false-belief responses, and do not increase young children's responding to above-chance levels. That such changes yielded chance rather than below-chance performance in young children, means that executive functioning difficulties may well be encouraging children to systematically err in reporting reality. Removing such task difficulties, however, does not reveal systematic correct performance for the youngest children.

Attempts to increase the salience of mental contents are captured in the meta-analysis with the variable salience. Manipulations under this variable, in particular, those that involve stating or picturing the protagonist's mental contents, do enhance performance. Performance is enhanced at all ages, however, and young children do not achieve above-chance performance. Hence, in neither of these cases do the findings conform with early competence accounts of false-belief performance.

Recently, Carlson and Moses (in press) conducted a thorough individual differences analysis of performance on theory-of-mind tasks (including, focally, several false-belief tasks) and executive functioning tasks (including, focally, inhibitory-control tasks). (See Hughes, 1998 for similar, but less comprehensive, results.) Inhibitory-control tasks correlated

highly with theory-of-mind tasks (approximately .60) and a significant correlation remained even after age, verbal intelligence, and several other control measures were partialled out. At the same time, however, a "theory-of-mind factor" emerged in a principal component analysis alongside executive functioning factors, and regression analyses indicated significant independent contributions of theory-of-mind and executive functioning to false-belief performance. Thus, Carlson and Moses (in press, p. 119, 122) concluded that "executive or inhibitory constraints alone would appear to underdetermine the nature of the changes taking place in young children's understanding of mind," and advocated an account "in which executive functioning interacts with conceptual difficulties" (see Perner & Lang, 1999, for similar conclusions).

The meta-analytic findings help to illuminate the potential interrelation of conceptual change and executive functioning. Using Figure 2B as a template, note that across age, false-belief conceptual understanding changes from being absent to present. At any one age, however, there is variation in performance across individuals and across tasks. These variations, in part, reflect inhibitory demands and inhibitory control. At the youngest ages, for example, children with better inhibitory control skills (or children on tasks that require less rather than more inhibitory control) perform better. Specifically, these youngest children seem to perform better by avoiding systematic errors, which makes their performance higher relative to same-age peers who systematically err. The absence of genuine conceptual understanding for such young children, however, means that performance does not increase to above-chance levels. At intermediate ages, better inhibitory control could also aid transitional children: a transitional child with poor inhibitory control might perform at chance, whereas a peer with more skill might perform at above chance. At each age (or with age partialled out), therefore, executive functioning could be highly correlated with performance. But across ages, performance still reveals genuine conceptual change.

Again, there may well be a deeper conceptual connection between theory of mind and executive functioning, beyond a methodological concern about false-belief tasks. Carlson and Moses (in press), for example, extended a distinction first advanced by Russell (1996), to distinguish between executive performance problems and executive construction problems. Performance problems are the type addressed thus far; perhaps immature executive functioning prohibits young children from performing correctly on standard false-belief tasks. Consider instead exec-

utive construction problems. Developmentally normative deficits in executive functioning may mean that young children rarely attend to the world of the mind over the prepotent world of reality. The salience of reality, together with poor skills at overcoming that salience, could lead young children to rarely consider the realm of mental contents. Only when inhibitory control skills are better developed can they suppress this reality orientation sufficiently to productively think about the mind. Under this proposal, then, executive functioning developments may be required for conceptual change in this domain, because they enable a specific conceptual orientation to develop.

This intriguing hypothesis must take its place beside several others that have been put forth to explain the forces that lead to conceptual change. For the present discussion, what is notable is that such hypotheses assume the presence of conceptual change. The meta-analysis provides support for this important assumption.

Conceptual Change Accounts

Conceptual change accounts come in several competing varieties. The meta-analysis by itself, however, provides little information for definitively choosing among alternative explanations for the underlying nature of conceptual change, because it does not compare false-belief performance with theoretically chosen contrasting conceptions and tasks. For example, returning to Figure 1, Wellman and colleagues (Bartsch & Wellman, 1995; Wellman & Woolley, 1990) proposed that young children's initial understanding of persons amounts to a desire psychology—a theory of persons based on an initial, simplified understanding of three internal states: states of emotion, states of perception, and, especially, states of desire. This understanding is conceptually quite different from adults' belief-desire psychology in that young children fail to understand persons as having internal mental representations of the world (prototypically beliefs), and thus fail to see mind and action as jointly determined by beliefs and desires. The meta-analytic findings are consistent with such an account, but tests of this account require, among other things, research directly comparing children's developing conceptions of desires as well as beliefs (e.g., Bartsch & Wellman, 1995).

The meta-analytic findings are consistent with alternative accounts as well. For instance, Perner (1991) focuses on young children's understanding of representations. He argues that young children's initial theory of mind is nonrepresentational, including only a simplified nonrepresentational understanding of

thinking, believing, or pretending. Additionally, and in a similar way, young children fail to understand physical representations, such as photographs and drawings, as genuinely representational. Only as children acquire an understanding of representations in general, do they come to understand beliefs and false beliefs. Indeed this understanding of representation changes children's understanding of desires, knowledge, memory, and more, as they now develop a generally representational theory of mind. Tests of this account have involved looking beyond an understanding of beliefs and false beliefs to children's developing conceptions of photographs, models, and other representations (e.g., Leekam & Perner, 1991; Zaitchik, 1990). Again, the meta-analytic finding that young children develop a genuine understanding of false beliefs (and hence mental misrepresentations of reality) is consistent with such an account, but is also consistent with competing conceptual change accounts (see, e.g., Slaughter, 1998). In general, the meta-analysis did not include consideration of the myriad comparison tasks needed for such theory-building efforts. Neither do many of the studies of children's understanding of false beliefs. We hope that one outcome of the meta-analysis will be to diminish research narrowly focused on false-belief performance alone and to facilitate efforts to understand development within children's theory of mind by focusing on patterns of performance across a theoretically well-chosen package of comparison tasks.

Yet, the meta-analysis did yield data specifically relevant to two issues within competing conceptual change accounts. The first concerns false-belief errors. Significantly below-chance performances have been proposed to be indicative of the nature of specific erroneous conceptions of belief. For example, consider Wellman's (1990) proposal that young children may have a "copy" understanding of beliefs—the conception that belief always veridically captures reality—and thus a person necessarily believes what is true. The meta-analytic data show, however, several conditions under which young children do not make systematic false-belief errors, but rather perform at chance level. It seems unlikely, therefore, that young children first hold a definite copy misconception about belief. It is more likely that they fail to understand belief altogether and understand human behavior instead via other constructs such as the person's desires, emotions, or perceptions.

Clarification is needed here, however. When the meta-analytic data show children performing at chance, this represents the mean of a group of children's judgments. Such a mean could result (1) from

individual children largely performing at chance, or (2) from half the children understanding false belief and judging correctly, and the other half making systematic errors. To differentiate those possibilities data are needed on individual children's patterns of response across multiple false-belief trials. Most studies, however, report responses for single trials or sum responses across children and trials. Nonetheless, the meta-analysis provides some indirect evidence that group means that are close to chance levels often represent confused, random performance at the individual level as well. Recall that data about consistency over trials were available for 52 conditions. Indeed, as reported earlier, consistency in terms of proportion of agreement is correlated linearly with age, $r(51) = .28$, $p < .05$. Thus, younger children's answers are more likely to be inconsistent across trials than are older children's. At the same time, as is clear from other analyses, younger children's answers are more likely to be at-chance rather than above-chance level. For these 52 conditions we have compared children's consistency and correctness more directly. In order to score a group's mean performance as close to or deviant from chance (.50 correct) the absolute value of the group mean minus .50 was calculated (yielding scores that ranged from 0 to .50, where 0 is chance performance and .50 is perfectly correct or perfectly incorrect). This deviation score correlates significantly with consistency, $r(51) = .34$, $p < .02$. This correlation means that, on average, group means that are close to chance are more likely to be inconsistent than consistent. Thus, mean scores that are at chance often include a sizable proportion of random, confused performance rather than even mixes of systematically correct versus incorrect responses.

A second focal issue for competing conceptual change accounts concerns the role of one's own mental states in the comprehension or attribution of others' mental states. The relation between understanding one's own mental states and understanding others' has been hotly debated by philosophers and psychologists at least since Descartes. Within the area of theory of mind, the debate is manifest in differences between simulation-theory accounts and theory-theory accounts. Simulation theorists argue that there is a special primacy to knowing one's own mental contents (e.g., Harris, 1992). First-person experience not only has an immediacy and vividness that informs an understanding of mind, but understanding other minds requires using one's own experience to simulate that of others. In contrast, theory-theorists stress the development of an interrelated body of knowledge, based on core mental-state constructs such as "beliefs" and "desires," that apply to all per-

sons generically, that is, to both self and others (Gopnik, 1993; Gopnik & Wellman, 1994).

The meta-analytic findings comparing judgment about the false beliefs of self versus others is of clear relevance here: simulation accounts emphasizing the primacy of self-experience suggest that self-understanding should develop first. As shown in Figure 5, however, performance on false-belief tasks for self and for others is virtually identical at all ages. The lack of differences between false-belief judgments for self and others has been reported in several individual studies (e.g., Gopnik & Astington, 1988). Yet relatively small sample sizes in these studies (e.g., 16–20 participants) mean that a theoretically important, albeit empirically small, difference could easily have gone undetected. The meta-analytic comparison, however, summarizes the performance of several thousand children.

On the surface, the fact that children ever systematically err in reporting their own false beliefs seems problematic for simulation accounts. Again, however, data from the meta-analysis, although important, are not definitive. For example, Harris (1992) has argued that young children's difficulties with false-belief tasks for self are memory difficulties. False-belief tasks for self require not a report of current mental states, but memory for past states. Reporting past mental states may be misleadingly difficult because children must overcome their own current (correct) belief and report their own prior (incorrect) one. Again, more detailed theoretical comparisons require consideration of other tasks beyond false belief tasks.

CONCLUSIONS

The current meta-analysis organizes the available findings on false-belief understanding—a sizable accomplishment, given that the voluminous accumulating findings had begun to seem contradictory and interpretively intractable. It is now clear that across studies, when organized systematically, the results are largely robust, orderly, and consistent. Theoretically, once the findings are clarified, several competing accounts of false-belief performance can be evaluated. In particular, early competence accounts that claim apparent developments during the ages of 3 to 5 years are solely the products of overly difficult tasks masking young children's essentially correct understanding of belief are not substantiated in several key regards.

The meta-analysis also argues against proposals that an understanding of belief, including false belief, is the culture-specific product of socialization within literate, individualistic Anglo-European cul-

tures (Lillard, 1998). A mentalistic understanding of persons that includes a sense of their internal representations—their beliefs—is widespread. Although children may acquire such conceptions sooner or later depending on the cultural communities and language systems in which they are reared, young children in Europe, North America, South America, East Asia, Australia, and Africa, and from nonschooled “traditional” as well as literate “modern” cultures, all acquire these insights on roughly the same developmental trajectory. Of course, a radically different pattern of results may yet be discovered in some presently untested sociocultural childhood milieu. Even if an understanding of actions in terms of beliefs proves to be not strictly universal, the meta-analysis documents that it is impressively widespread, at least in childhood. This suggests that such a conception is a natural, easily adopted way of understanding persons worldwide; it is cognitively “contagious,” to use Sperber’s (1990) terminology.

Methodologically, these results inform us about several task variations that are essentially equivalent. This frees investigators to use a specific task instantiation because it is easier for them to administer than others (e.g., using puppets versus people as the target character) or because it best fits their theoretical purposes (e.g., for some purposes an unexpected-contents task may most closely parallel a target-contrast task, whereas for other purposes a change-of-locations task may do so). At the same time, the meta-analytic results show that some forms of the tasks do enhance children’s performance. Any investigator who is interested in assessing younger children’s first emerging understanding of belief would do well to consider tasks that are framed in terms of deception, engage the child in actively transforming the situation, and have no real and present object available at the time the target question is asked.

Note, however, that when a task variation increases performance, for example, from below-chance to chance levels or from chance to above-chance levels, this increased performance can be interpreted in one of two ways: The manipulation may have resulted in a better, more sensitive test of young children’s understanding, or it may have resulted in an artifactually easy task that is prone to false positives. Consider, for example, the earlier discussion of Sheffield et al.’s (1993) simplified task. Moreover, some of the controversies surrounding deception concern whether deception methods more sensitively test false-belief understanding (as argued, e.g., by Chandler & Hala, 1994; Sullivan & Winner, 1993) or merely increase false positives (as argued, e.g., by Sodian, 1994). The meta-analysis cannot definitively address

this issue. It is important to note, however, that the main meta-analytic findings hold regardless. That is, when the data from many studies are pooled, such task changes (regardless of their ultimate interpretation) fail to influence the underlying developmental trajectory; with increasing age, children’s judgments proceed from incorrect to significantly correct performance, for more difficult tasks as well as easier tasks.

Equally important methodologically, the meta-analysis underwrites the increasingly common practice of using false-belief tasks as a marker of theory-of-mind understanding in other types of research, such as research on individual differences in early social cognition (e.g., Hughes & Dunn, 1998) and on impaired or delayed social cognition with autistic individuals (e.g., Baron-Cohen, 1995). False-belief tasks demonstrably provide a robust measure of an important early development. Nevertheless, several cautions are necessary. (1) False-belief tasks measure only one narrow aspect of social cognitive development; therefore, use of a battery of social cognition tasks would be best for many studies. (2) Individual differences can take several forms: Some individual difference dimensions (perhaps shyness) can differentiate individuals across the lifespan; other differences emerge only within a particular window of development and thus reflect an individual’s speed of attaining a common developmental milestone. False-belief measures fall into the latter category, and their nature needs to be carefully considered in research on individual differences. (3) To be useful in individual-differences research a measure needs to be a valid representative of the proposed construct, and must have certain psychometric properties. Critically, reliability and unreliability of measurement place limits on the assessment of interrelation among variables of interest. The meta-analysis suggests that, within its developmental window of usefulness, false-belief measures are reliable. To reiterate, the meta-analysis includes 52 reports of children’s consistency of responding across equivalent false-belief tasks. On average, consistency, or proportion of agreement, was .84; that is, 84% of the time children’s first false-belief response was matched in their second response.

Investigators have seldom explicitly considered reliability of false-belief performances. One exception is a study by Mayes, Klin, Terycak, Cicchetti, and Cohen (1996) who gave 23 children three false-belief tasks in one session and then the same tasks in a second session 2 to 3 weeks later. Data within a session were similar to the consistency data reported from the meta-analysis. For example, within

their second session, 11 children were incorrect and 7 correct for all three false-belief tasks (with the remaining 5 answering one or two tasks correctly). Hence, proportion of agreement was .78, even when considered across all three tasks.

Mayes et al. (1996) assessed test-retest reliability by comparing responses to identical tasks in Session 1 versus 2. Proportion of agreement for false-belief answers ranged from .58 to 1.0 and averaged .75. However, Mayes et al. had more complete data and were able to calculate kappa, an index that takes into account not only proportion of agreement but also a baseline of expected agreement on the basis of chance alone. Only three of eight test-retest comparisons achieved "acceptable" kappas (.70 or greater). How best to interpret these results is unclear. The majority of test-retest responses that failed to agree were cases of improvement—initially incorrect children improved 3 weeks later. Improvements may represent legitimate developments rather than simple statistical disagreement. Moreover, with only 23 subjects, chance estimates for calculating kappas are noisy at best. Finally, the authors only reported test-retest reliability across pairs of identical tasks. This seems to neglect information available from the high proportions of agreement within a session. For example, suppose a child was considered as understanding false belief only if correct on two of three or more tasks in Session 1. What would the test-retest reliability be of such a composite score, comparing scores from session 1 to Session 2? In total, the Mayes et al. (1996) study is consistent with the meta-analysis in demonstrating high reliability for false-belief judgments within sessions. The question of longer term test-retest reliability remains an open one for further research.

Finally, under the heading of methodological implications, it seems to us that the meta-analysis should lay to rest a great many questions about how task modifications enhance performance, thereby reducing the volume of studies designed simply to assess such questions. In general, the accumulation of 591 conditions includes task variations of sufficient scope to obviate the need for still one more variation. There will always be defensible exceptions to this caveat. For example, it might be of some use to empirically assess, within the scope of a single study, the effects of a false-belief task that includes all four performance-enhancing features outlined earlier. For the most part, however, researchers should turn to looking at the important theoretical questions that are outstanding, such as what mechanisms account for the developmental changes now clearly evident in children's understanding of belief and of

the mind. The results of the meta-analysis, summarizing 591 false-belief conditions, can facilitate such theoretically driven research on still broader issues in the field.

ACKNOWLEDGMENTS

This research was supported by a grant from NICHD (HD-22149) to the first author. The authors thank the many investigators who sent us amplified descriptions of their data or methods, and Janet Astington who suggested the subtitle.

ADDRESSES AND AFFILIATIONS

Corresponding author: Henry M. Wellman, Center for Human Growth and Development, University of Michigan, 300 N. Ingalls, 10th Floor, Ann Arbor, MI 48109-0406; e-mail: hmw@umich.edu. David Cross is at Texas Christian University, Fort Worth. Julianne Watson is also at the University of Michigan.

REFERENCES

- Astington, J. W. (1993). *The child's discovery of the mind*. Cambridge, MA: Harvard University Press.
- Astington, J. W., Gopnik, A., & O'Neill, D. (1989). *Young children's understanding of unfulfilled desire and false belief*. Unpublished manuscript, Ontario Institute for Studies in Education, Toronto, Ontario, Canada.
- 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.
- Avis, J., & Harris, P. L. (1991). Belief-desire reasoning among Baka Children. *Child Development*, 62, 460–467.
- Baron-Cohen, S. (1991). Do people with autism understand what causes emotion? *Child Development*, 62, 835–895.
- Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind?" *Cognition*, 21, 37–46.
- Bartsch, K. (1996). Between desires and beliefs: Young children's action predictions. *Child Development*, 67, 1671–1685.
- Bartsch, K., London, K., & Knowlton, D. (1997, April). *Children's use of belief information in selecting hypothetical persuasive strategies*. Poster session presented at the biennial meeting of the Society for Research in Child Development, Washington, DC.
- Bartsch, K., & Wellman, H. M. (1989). Young children's attribution of action to beliefs and desires. *Child Development*, 60, 946–964.
- Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York: Oxford University Press.
- Berguno, G. (1997). *Young children's understanding of representations do not provide them with a theory of mind*. Unpublished manuscript, City University, London, U.K.

- Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development, 70*, 381–395.
- Carlson, S. M., & Moses, L. (in press). Individual differences in inhibitory control and children's theory of mind. *Child Development*.
- Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development, 69*, 672–691.
- Carpendale, J. I., & Chandler, M. J. (1996). On the distinction between false belief understanding and subscribing to an interpretive theory of mind. *Child Development, 67*, 1686–1706.
- Chandler, M. (1988). Doubt and developing theories of mind. In J. Astington, P. Harris, & D. Olson (Eds.), *Developing theories of mind*. New York: Cambridge University Press.
- Chandler, M., Fritz, A. S., & Hala, S. (1989). Small scale deceit: Deception as a marker of 2-, 3-, and 4-year-olds early theories of mind. *Child Development, 60*, 1263–1277.
- Chandler, M., & Hala, S. (1994). The role of personal involvement in the assessment of early false belief skills. In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind* (pp. 403–425). Hillsdale, NJ: Erlbaum.
- Chen, M. J., & Lin, Z. X. (1994). Chinese preschoolers' difficulty with theory-of-mind tests. *Bulletin of the Hong Kong Psychological Society, 32/33*, 34–46.
- Clements, W., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*, 377–397.
- Custer, W. L. (1996). A comparison of young children's understanding of contradictory representations in pretense, memory, and belief. *Child Development, 67*, 678–688.
- Dalke, D. E. (1995). Explaining young children's difficulty on the false belief task: Representational deficits or context-sensitive knowledge? *British Journal of Developmental Psychology, 13*, 209–222.
- Davis, D. L. (1997, April). *Children's understanding of the role of knowledge and thinking in pretense*. Paper presented at the biennial meeting of the Society for Research in Child Development, Washington, DC.
- Dennett, D. C. (1979). *Brainstorms*. Hassocks, Sussex: Harvester.
- Draper, N. R., & Smith, H. (1981). *Applied regression analyses* (2nd ed.). New York: Wiley.
- 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.
- Flavell, J. H. (1988). The development of children's knowledge about the mind: From cognitive connections to mental representations. In J. Astington, P. Harris, & D. Olson (Eds.), *Developing theories of mind* (pp. 244–267). New York: Cambridge University Press.
- 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 (Vol. Eds.), W. Damon (Series Ed.) *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (pp. 851–898). New York: Wiley.
- Flavell, J. H., Mumme, D. L., Green, F. L., & Flavell, E. R. (1992). Young children's understanding of different types of beliefs. *Child Development, 63*, 960–977.
- Fodor, J. A. (1992). A theory of the child's theory of mind. *Cognition, 44*, 283–296.
- Freedman, D. A., & Peters, C. (1984). Bootstrapping a regression equation: Some empirical results. *Journal of the American Statistical Association, 79*, 97–106.
- Freeman, N. H., & Lacohee, H. (1995). Making explicit 3-year-olds' implicit competence with their own false beliefs. *Cognition, 56*, 31–60.
- Freeman, N. H., Lewis, C., & Doherty, M. J. (1991). Preschoolers's grasp of a desire for knowledge in false-belief prediction. *British Journal of Developmental Psychology, 9*, 139–157.
- Fritz, A. S. (1992). *Event saliency as a constraint upon young children's developing theories of mind*. Unpublished doctoral dissertation, University of British Columbia, Vancouver, British Columbia, Canada.
- Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development, 10*, 483–527.
- Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3–7 years old on a stroop-like day–night test. *Cognition, 53*, 129–153.
- Ghim, H.-R. (1997, April). *Three-year-olds' failure in the standard change-of-location false belief task*. Paper presented at the biennial meeting of the Society for Research in Child Development, Washington, DC.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Gopnik, A. (1993). How we know our minds: The illusions of first person knowledge of intentionality. *Behavioral and Brain Sciences, 16*, 1–14.
- Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance–reality distinction. *Child Development, 59*, 26–37.
- Gopnik, A., & Slaughter, V. (1991). Young children's understanding of changes in their mental states. *Child Development, 62*, 98–110.
- Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. Hirschfeld & S. Gelman (Eds.), *Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press.
- Green, B. F., & Hall, J. A. (1984). Quantitative methods for literature reviews. *Annual Review of Psychology, 35*, 37–53.
- Hala, S., & Chandler, M. (1996). The role of strategic planning in accessing false-belief understanding. *Child Development, 67*, 2948–2966.
- Hala, S., Chandler, M., & Fritz, A. S. (1991). Fledgling theories of mind: Deception as a marker of 3-year-olds' understanding of false belief. *Child Development, 62*, 83–97.
- Happe, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*, 843–855.
- Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind and Language, 7*, 120–144.

- Harris, P. L., Johnson, C. N., Hutton, D., Andrews, G., & Cooke, T. (1989). Young children's theory of mind and emotion. *Cognition and Emotion*, 3, 379-400.
- Hickling, A. K., Wellman, H. M., & Gottfried, G. (1997). Preschoolers' understanding of others' mental attitudes toward pretend happenings. *British Journal of Developmental Psychology*, 15, 339-354.
- 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.
- Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.
- Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology*, 16, 233-253.
- Hughes, C., & Dunn, J. (1998). Understanding and emotion: Longitudinal associations with mental-state talk between young friends. *Developmental Psychology*, 34, 1026-1037.
- Johnson, C. N., & Maratsos, M. P. (1977). Early comprehension of mental verbs: Think and know. *Child Development*, 48, 1743-1747.
- Kikuno, H. (1997, April). *Recall and construction of false belief: Failure arising from processing deficiency*. Paper presented at the biennial meeting of the Society for Research in Child Development, Washington, DC.
- Kochanska, G., Murray, K., Jacques, T. Y., Koenig, A. L., & Vandergeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development*, 67, 490-507.
- Koyasu, M. (1996). Understanding representations of self, other person, and photograph in young children. In *The emergence of human cognition and language*, vol. 3. A report to the Ministry of Education, Science, Sports and Culture. Tokyo: Japan.
- 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.
- Leekam, S., & Perner, J. (1991). Does the autistic child have a "metarepresentational" deficit? *Cognition*, 40, 203-218.
- Leslie, A. M. (2000). How to acquire a "representational theory of mind." In D. Sperber & S. Davis (Eds.), *Meta-representation* (pp. 197-223). Oxford, U.K.: Oxford University Press.
- Leslie, A. M., & Roth, D. (1993). What autism teaches us about meta representation. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from autism* (pp. 83-111). Oxford, U.K.: Oxford University Press.
- Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, 43, 225-251.
- Lewis, C., Freeman, N. H., Hagestadt, E., & Douglas, H. (1994). Narrative access and production in preschoolers' false belief reasoning. *Cognitive Development*, 9, 397-424.
- Lewis, C., & Osborne, A. (1990). Three-year-olds' problems with false belief: Conceptual deficit or linguistic artifact? *Child Development*, 61, 1514-1519.
- Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin*, 123, 3-32.
- Lillard, A. S., & Flavell, J. H. (1992). Young children's understanding of different mental states. *Developmental Psychology*, 28, 626-634.
- Mayes, L. C., Klin, A., & Cohen, D. J. (1994). The effect of humour on children's developing theory of mind. *British Journal of Developmental Psychology*, 12, 555-561.
- Mayes, L. C., Klin, A., Tercyak, K. P., Cicchetti, D. V., & Cohen, D. J. (1996). Test-retest reliability for false-belief tasks. *Journal of Child Psychology and Psychiatry*, 37, 313-319.
- Mazzoni, G. (1995). *Is there any understanding of mental state in 2-year-olds? Some evidence from deception*. Unpublished manuscript, Università di Firenze, Italy.
- Miller, S. A. (in press). Children's understanding of preexisting differences in knowledge and belief. *Developmental Review*.
- Mitchell, P. (1996). *Acquiring a conception of mind: A review of psychological research and theory*. Hove, U.K.: Psychology Press.
- Mitchell, P., & Lacohee, H. (1991). Children's early understanding of false belief. *Cognition*, 39, 107-127.
- Moore, C., Furrow, D., Chiasson, L., & Patriquin, M. (1994). Developmental relationships between production and comprehension of mental terms. *First Language*, 14, 1-17.
- Moore, C., Pure, K., & Furrow, P. (1990). Children's understanding of the modal expression of certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development*, 61, 722-730.
- Moses, L. J. (1993). Young children's understanding of belief constraints on intention. *Cognitive Development*, 8, 1-25.
- Moses, L. J., & Flavell, J. H. (1990). Inferring false beliefs from actions and reactions. *Child Development*, 61, 929-945.
- Naito, M., Komatsu, S., & Fuke, T. (1994). Normal and autistic children's understanding of their own and others' false belief: A study from Japan. *British Journal of Developmental Psychology*, 12, 403-416.
- Nelson, K., Plesa, D., & Henseler, S. (1998). Children's theory of mind: An experiential interpretation. *Human Development*, 41, 7-29.
- Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.
- Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences*, 3, 337-344.
- 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., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: You catch it from your sibs. *Child Development*, 65, 1228-1238.
- Perner, J., & Wimmer, H. (1987). Young children's understanding of belief and communicative intention. *Pakistan Journal of Psychological Research*, 2, 17-40.
- Peterson, C. C., & Siegal, M. (1995). Deafness, conversation and theory of mind. *Journal of Child Psychology and Psychiatry*, 36, 459-474.

- Phillips, W. (1994). *Understanding intention and desire by children with autism*. Unpublished doctoral dissertation, University of London, London, U.K.
- Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, 13, 73–90.
- Robinson, E. J., & Mitchell, P. (1992). Children's interpretation of messages from a speaker with a false belief. *Child Development*, 62, 639–652.
- Robinson, E. J., & Mitchell, P. (1994). Young children's false-belief reasoning: Interpretation of messages is no easier than the classic task. *Developmental Psychology*, 30, 67–72.
- Robinson, E. J., & Mitchell, P. (1995). Masking children's early understanding of the representational mind: Backwards explanation versus prediction. *Child Development*, 66, 1022–1039.
- Robinson, E., Riggs, K., & Peterson, D. (1997, April). *Are realist errors in false belief tasks due to difficulty with counterfactual reasoning?* Paper presented at the biennial meeting of the Society for Research in Child Development, Washington, DC.
- Robinson, E. J., Riggs, K. J., & Samuel, J. (1996). Children's memory for drawings based on a false belief. *Developmental Psychology*, 32, 1056–1064.
- Roth, D., & Leslie, A. (1998). Solving belief problems: Toward a task analysis. *Cognition*, 66, 1–31.
- Ruffman, T., Olson, D. R., Ash, T., & Keenan, T. (1993). The ABCs of deception: Do young children understand deception in the same way as adults? *Developmental Psychology*, 29, 74–87.
- Russell, J. (1996). *Agency: Its role in mental development*. Hove, U.K.: Erlbaum.
- Russell, J., & Jarrold, C. (1994). *Executive factors in the false belief task, and prospects for an executive account of theory of mind development*. Unpublished manuscript, University of Cambridge, Cambridge, U.K.
- Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The 'windows task' as a measure of strategic deception in preschoolers and autistic subjects. *British Journal of Developmental Psychology*, 9, 331–349.
- Saltmarsh, R., Mitchell, P., & Robinson, E. (1995). Realism and children's early grasp of mental representation: Belief-based judgements in the state change task. *Cognition*, 57, 297–325.
- Seier, W. L. (1993, March). *A comparison of young children's understanding of contradictory mental representations in pretense, memory, and belief*. Poster session presented at the biennial meeting of the Society for Research in Child Development, New Orleans, LA.
- Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of first reference to mental state. *Cognition*, 14, 301–321.
- Sheffield, E. G., Sosa, B. B., & Hudson, J. A. (1993, March). *Understanding metarepresentation: 2- and 3-year-olds' comprehension of false belief*. Paper presented at the biennial meeting of the Society for Research in Child Development, New Orleans, LA.
- Siegal, M. (1997). *Knowing children: Experiments in conversation and cognition* (2nd ed.). Hove, U.K.: Erlbaum.
- Siegal, M., & Beattie, K. (1991). Where to look first for children's understanding of false beliefs. *Cognition*, 38, 1–12.
- Slaughter, V. (1998). Children's understanding of pictorial and mental representations. *Child Development*, 69, 321–332.
- Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: Training children to understand belief. *Child Development*, 67, 2967–2988.
- Sodian, B. (1994). Early deception and the conceptual continuity claim. In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind*. Hove, U.K.: Erlbaum.
- Sodian, B., Taylor, C., Harris, P. L., & Perner, J. (1991). Early deception and the child's theory of mind. *Child Development*, 62, 468–483.
- Spelke, E. S. (1994). Initial knowledge: Six suggestions. *Cognition*, 50, 431–445.
- Sperber, D. (1990). The epidemiology of beliefs. In C. Fraser & G. Gaskell (Eds.), *The social psychological study of widespread beliefs*. Oxford: Clarendon.
- 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-olds' understanding of mental states: The influence of trickery. *Journal of Experimental Child Psychology*, 56, 135–148.
- Taylor, M. (1996). A theory of mind perspective on social cognitive development. In R. Gelman & T. Au (Eds.), *Handbook of perception and cognition: Vol. 13. Perceptual and cognitive development* (pp. 283–329). New York: Academic.
- Taylor, M., & Carlson, S. M. (1997). The relation between individual differences in fantasy and theory of mind. *Child Development*, 68, 436–455.
- Vinden, P. G. (1996). Junin Quechua children's understanding of mind. *Child Development*, 67, 1707–1716.
- Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press, A Bradford Book.
- Wellman, H. M., & Bartsch, K. (1988). Young children's reasoning about beliefs. *Cognition*, 30, 239–277.
- Wellman, H. M., & Gelman, S. A. (1998). Knowledge acquisition in foundational domains. In D. Kuhn & R. Siegler (Vol. Eds.), N. Eisenberg (Series Ed.) *Handbook of Child Psychology: Cognition, perception, and language* (5th ed., pp. 523–573). New York: Wiley.
- 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., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, 35, 245–275.
- Wimmer, H., & Hartl, M. (1991). Against the Cartesian view on mind: Young children's difficulty with own false beliefs. *British Journal of Developmental Psychology*, 9, 125–128.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–128.
- Wimmer, H., & Weichbold, V. (1994). Children's theory of mind: Fodor's heuristics examined. *Cognition*, 53, 45–57.
- Winner, E., & Sullivan, K. (1993). *Deception as a zone of prox-*

- imal development for false belief understanding.* Unpublished manuscript, Boston College, Boston, MA.
- Woolley, J. D. (1995). Young children's understanding of fictional versus epistemic mental representations: Imagination and belief. *Child Development, 66*, 1011–1021.
- Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin, 124*, 283–307.
- Yoon, M. G., & Yoon, M. (1993). *Development of children's understanding of false beliefs: Prediction of action precedes explanation of prediction, attribution of thinking, or retroactive justification.* Unpublished manuscript, Dalhousie University, Halifax, Nova Scotia.
- Youngblade, L. M., & Dunn, J. (1995). Individual differences in young children's pretend play with mother and sibling: Links to relationships and understanding of other people's feelings and beliefs. *Child Development, 66*, 1472–1492.
- Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition, 35*, 41–68.
- Zaitchik, D. (1991). Is only seeing really believing? Sources of true belief in the false belief task. *Cognitive Development, 6*, 91–103.
- Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of General Psychology, 1*, 1–29.