JOURNAL OF COGNITION AND DEVELOPMENT, *7*(2), 139–172

Theories of Mind in Transition: A Microgenetic Study of the Development of False Belief Understanding

Jennifer Amsterlaw and Henry M. Wellman
University of Michigan

Microgenetic methods were used to document young children's (*N* = 36; *M* age = 3;5) acquisition of false belief (FB) understanding and investigate developmental mechanisms. A control group received no experience with FB; 2 other groups received microgenetic sessions designed to promote FB understanding. Over consecutive weeks, microgenetic groups received implicit feedback about their performance on 24 FB tasks and generated explanations for FB events. Only 1 microgenetic group improved. Differences in the schedules of microgenetic experience and in the amount and type of FB explanation children engaged in accounted for these differences. Improving children developed FB understandings gradually and exhibited fluctuating task performance, suggesting slow conceptual restructuring, not sudden insight. This work provides the first microgenetic record of children's transition to a representational theory of mind.

Between 3 and 5 years, children achieve one of the hallmarks of an adult-like theory of mind—an understanding of mental representation (see Flavell & Miller, 1998; Taylor, 1996; Wellman, 2002). This development is indexed in part by older children's success on false belief (FB) tasks that involve predicting the thoughts or actions of someone whose beliefs about the world are mistaken. In *Locations* tasks (e.g., Wimmer & Perner, 1983), children predict where a character will search for an object whose location is changed during his absence. Older children (4- and 5-year-olds) correctly report that the character will search in the object's original location, but younger children (3-year-olds) indicate the current location. Similar

Correspondence should be addressed to Jennifer Amsterlaw, Institute for Learning and Brain Sciences, University of Washington, Box 357988, Seattle, WA 98195. E-mail: amsterj@u.washington.edu

results are found with *Contents* tasks (e.g., Perner, Leekam, & Wimmer, 1987), where children are shown a familiar container (e.g., candy box) that holds unexpected contents (e.g., pencils) and are asked to predict what a naive observer will think is inside. Older children report the canonical contents, but younger children report the current contents, apparently failing to grasp that the observer will be tricked. These basic age effects have been replicated numerous times, including cross-linguistically and cross-culturally, and are robust across countless variations of the tasks (Wellman, Cross, & Watson, 2001). Other tasks requiring an understanding of mental representation show similar age trends: Younger preschool children perform poorly in judgments about the sources of their knowledge (O'Neill & Gopnik, 1991), have difficulty reporting on their own prior false beliefs (Gopnik & Astington, 1988), and fail to distinguish between appearance and reality (Flavell, Green, & Flavell, 1986).

Children's transition to a representational theory of mind is thus well-documented empirically. However, much of the research that demonstrates it is cross-sectional, yielding clear "before" and "after" snapshots of development but failing to describe change processes directly. This approach leaves unanswered important questions about the nature of underlying change processes in theory of mind development, including questions about the path, rate, and breadth of change, its sources and mechanisms, and individual differences in change processes (Siegler, 1995). With regard to the path of change, for example, do children progress through distinct transitional periods in acquiring an understanding of FB? Wellman et al.'s (2001) earlier meta-analysis suggested that as children move from consistent failure to consistent success on FB tasks—roughly between the ages of 3½ and 4 years—they go through an identifiable period of confused, at-chance task performance. However, because these results were derived from cross-sectional evidence, it is unclear whether this pattern of change accurately describes the trajectory of FB understanding in individual children. Similarly, with regard to the rate and breadth of change, do children acquire an understanding of FB somewhat suddenly, as in a flash of insight, or is the concept worked out more gradually over time? Is newly acquired FB understanding shown across a variety of situations, or is it first restricted to a smaller set of contexts and only later generalized? Such questions get at the heart of what children's changing competence in theory of mind actually represents (e.g., whether it reflects conceptual restructuring or incremental learning) but go beyond the scope of prior data.

One way to address questions about processes of developmental change is with microgenetic methods, which provide especially "thick" descriptions of change through extended, closely spaced, longitudinal assessments. The basic approach is to choose a task representative of the cognition in question, hypothesize the types of everyday experiences that lead to change, and then provide a higher concentration of these experiences than ordinary. Observations of changing performance are then analyzed intensively to examine the processes giving rise to them (see Miller

& Coyle, 1999; Siegler & Crowley, 1991). Microgenetic methods have been used successfully to inform our understanding of conceptual change in a variety of domains (for reviews, see Kuhn, 1995; Miller & Coyle, 1999) and they could be particularly revealing for the study of theory of mind as well. Thus, our goal was to use microgenetic methods to track ongoing theory of mind development in a group of young preschoolers to identify critical features and mechanisms of their transition to an understanding of FB.

An essential characteristic of good microgenetic research is that some kind of relevant change must occur. Ideally, the goal is to document children's initial discovery of a new strategy or concept and then scrutinize the processes leading up to and ensuing from that discovery. Therefore, an optimal microgenetic environment is one that both regularly assesses children's changing competence on a given task and also provides experiences likely to foster the changes of interest. In prior microgenetic research, two kinds of experiences have been shown to foster change: (a) providing children with online feedback about their task performance and (b) asking them to explain correct solutions. For example, Siegler (1976) gave children Piagetian balance-scale problems and asked them to predict which way the balance would tip; when children were able to view the actual outcome following their predictions, their reasoning improved. In Siegler's (1995) microgenetic study of number conservation, asking children to explain the experimenter's correct solutions was even more effective than feedback alone in eliciting higher level reasoning. Our microgenetic study of theory of mind development made use of both feedback and explanation in the context of standard FB task scenarios. Over the course of 6 to 9 weeks, children received an extended series of FB task sessions that both assessed their current level of understanding and gave them regular opportunities: (a) to observe the outcomes of the FB scenarios they made predictions about, and thereby receive "implicit" feedback about the accuracy of those predictions, and (b) to attempt explanations for the actual, correct outcomes.

At least as important as microgenetic findings from other domains, there is preliminary evidence that such experiences, particularly those with explanation, may be influential for theory of mind development. For example, in home observations of family talk about internal states, Dunn and colleagues (Dunn, Brown, & Beardsall, 1991; Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991) showed that the frequency with which 2- and 3-year-old children provided explanations about the causes and consequences of everyday human events predicted their FB knowledge and perspective-taking abilities months later. Bartsch and Wellman (1995) examined children's and parents' early mental state talk for predictors of children's first genuine reference to belief and found that the strongest predictor by far was children's earlier production of psychological explanations (r = .84). Moreover, theoretically, "theory–theory" accounts of developmental change posit that children's knowledge about the mind advances when they encounter evidence in their daily lives that either builds on their current knowledge system, or reveals limitations in its explanatory or predictive powers (e.g., Gopnik & Wellman, 1994). Thus, observing events that confound one's expectations and attempting to explain them may be central mechanisms of development for theory of mind.

Our microgenetic approach shares certain similarities with various training approaches to theory of mind (e.g., see Lohmann & Tomasello, 2003). Both microgenetic and training methods are aimed at examining the effects of providing children with new experiences thought to provoke developmental change, and thus both approaches hold the potential to inform us about sources of such change. Training studies have indeed generated important new knowledge about the conditions under which children can—and cannot—be taught new understandings of the mind. For example, training work has underscored the role of social interaction and talk about mental states in children's developing competence in this area (e.g., Appleton & Reddy, 1996; Lohmann & Tomasello, 2003; see Discussion for further review). However, training studies differ from a microgenetic approach in two important ways. First, many training studies are focused on whether young children can be taught new theories of mind *at all*, and so they investigate whether children's knowledge can be improved under especially auspicious conditions, such as with direct teaching and explicit corrective feedback. Our microgenetic approach is more interested in evaluating how children might construct more sophisticated understandings of mind *without* explicit instruction or correction—conditions we feel more appropriately characterize the natural context in which these skills develop. It is clear from natural language research on parent–child talk about mental states that although such contexts provide children with important conceptual information about mental life, such "teaching moments" do not typically involve explicit correction of children's misconceptions nor direct teaching of correct ideas (see, e.g., Bartsch & Wellman, 1995; Sabbagh & Callanan, 1998).

Second, prior training studies have relied on pretest, intervention, and posttest designs and have not attempted to track developmental change over multiple intervening timepoints. This yields basic information about the conditions under which change is provoked, but unlike a microgenetic design, it does not provide a detailed description of how change occurs between pre- and posttest. In addition, most training studies assess change over relatively short periods of time (several days, or at most 2–3 weeks), whereas this microgenetic study was designed to describe changes unfolding over 6 to 9 weeks time.

In addition to this study, we know of two other studies also taking a microgenetic approach to studying theory of mind development (Flynn, O'Malley, & Wood, 2004; Wahl, 2001). Flynn et al. (2004) examined relations between 3-year-olds' FB understanding and their inhibitory control skills over a series of six longitudinal testing sessions. Wahl (2001) observed 3-year-olds' performance on FB and appearance–reality tasks over the course of 10 sessions, during which children received feedback about their performance on FB tasks but did not generate FB explanations (see Discussion for further description). In contrast to our own results, however, neither of these microgenetic studies saw reliable improvement in children's FB task performance to above-chance levels during testing. In addition, we report results from two separate microgenetic conditions that had very different outcomes. These factors allow us to probe questions about the nature and mechanisms of theory of mind development in some detail.

METHOD

Design

The basic design called for two groups: (a) a microgenetic group receiving rich exposure to experiences hypothesized to promote theory of mind development and showing improvement in FB understanding over time and (b) a time-lag control group receiving no special experiences with FB to document developments normally occurring in the same amount of time. However, in addition to these two groups, we also include results from another microgenetic condition, an initial condition that failed to promote development in children's theories of mind. Having detailed data on children's performance in two different microgenetic environments gives us additional leverage to address questions about the sources and processes of developmental change. Thus we report results from (a) a *focal microgenetic group*, which received key microgenetic experiences that successfully promoted development in theory of mind; (b) a *comparison microgenetic group*, which received a somewhat different set of FB experiences and did not show change; and (c) a time-lag *control group*. For convenience, we refer to these as the microgenetic, comparison, and control groups, respectively.

Participants

In total, 73 children between the ages of 3;2 (3 years, 2 months) and 4;6 from two preschools in a small Midwestern U.S. city were pretested. Of these, 24 children (33%) failed to meet inclusion criteria by passing more than one of the three standard FB tasks at pretest (21 children) or by failing all the control questions for the standard tasks, indicating comprehension problems (3 children). Subsequently, 2 children were excluded due to parental refusal for further testing, 6 children voluntarily ended participation, and 5 children were unable (e.g., due to prolonged absences) to participate further. The final sample consisted of 36 children, 12 children in each group: 4 boys and 8 girls in the microgenetic group (*M* age = 42.5 months, *SD* = 3.6), 5 boys and 7 girls in the comparison group (*M* age = 42.0 months, *SD* = 2.6), and 6 boys and 6 girls in the control group (*M* age = 42.6 months, *SD* = 2.8). Children were randomly assigned to microgenetic and control conditions, but the microgenetic group was run after we had already completed the comparison group. The sample was approximately 56% White, 34% Asian, and 10% other backgrounds. No measures of socioeconomic status were directly obtained, but the preschool populations were generally upper-middle class.

Materials

Pre- and posttest tasks included three standard FB measures: (a) a Locations task, (b) an "Other" version of the Contents task (where children predict another person's FB), and (c) a "Self" version of the Contents task (where children report their own prior FB). Also included were a FB Explanation task and a Seeing–Knowing task. Two additional tasks, an Appearance–Reality task and a FB Emotion task, were included in the posttest. Materials for the tasks consisted of small stuffed animals and plastic figurines, several small boxes and containers, and a variety of objects that could be placed inside the containers (keys, marbles, bandaids, etc.). Posttest materials were similar to those from the pretest but used different characters, containers, and objects. The Appearance–Reality task used a sponge painted to look like a rock, a white plastic egg and an 8.5 × 11-in. sheet of transparent blue plastic.

Materials for microgenetic sessions with the microgenetic and comparison groups consisted of hand-drawn storybooks and prop sets that were used to act out Locations and Contents tasks (see Table 1 for task descriptions). Wellman et al.'s (2001) meta-analysis shows that children perform similarly on FB tasks regardless of format (e.g., toy figures, stories, real people); we used two different formats to maintain children's interest over the 24 tasks.

Procedures

Children were tested individually in a private room at their preschool. For each group, two trained female researchers conducted the testing. Most children saw the same researcher for the pretest and all of their microgenetic sessions, but not the posttest. Posttesting in these cases was conducted by a researcher who was also familiar to children because of time spent in their classroom. Because of scheduling constraints, it was not possible to have separate researchers conduct posttesting for all the children; six children in the microgenetic group, four children in the comparison group, and four children in the control group saw the same researcher for all of their sessions. All sessions were videotaped.

Pretest. For the Locations task, children predicted where a toy bear would look for his penny after it was moved from its original location (box) to a new one (bag) while he napped. In the Contents–Other task, children predicted what a naive doll would think was inside a raisin box that actually contained a marble. In the Contents–Self task, children were asked to report their previous FB about the con-

TABLE 1
False Belief Tasks Used in Microgenetic Sessions

Locations Tasks[a]		*Contents Tasks*[b]	
1. Marcia's muffins (prop)	Muffins moved from drawer to fridge	1. Playdoh can (prop)	Playdoh can has ball inside
2. Frogatha's fly (prop)	Fly moved from sand pile to lily pad	2. Candy (prop)	Candy box has buttons inside
3. Samantha's cookie (prop)	Cookie moved from oven to fridge	3. Cereal box (prop)	Cereal box has candle inside
4. Geoffrey's butterfly (prop)	Butterfly goes from net to flower	4. Graham crackers (prop)	Cracker box has small book inside
5. Tigger's cow (prop)	Cow goes from barn to haystack	5. Legos (prop)	Legos box has crayons inside
6. Frannie's picture (prop)	Picture moved from under bed to desk	6. Lunchbox (book)	Lunchbox has doll inside
7. Billy's rabbit (book)	Rabbit goes from rabbit house to under bed	7. Honey pot (book)	Honey pot has carrots inside
8. Fifi's bone (book)	Bone moved from dog house to watering can	8. Mailbox (book)	Mailbox has apple inside
9. Ernie's rubber duckie (book)	Rubber duckie moved from bathtub to cupboard	9. Dogfood (book)	Dogfood box has soap inside
10. Simon's turtle (book)	Turtle moved from bucket to garden	10. Cookie jar (book)	Cookie jar has flower inside
11. Spiffer's clothes (book)	Clothes moved from flower to mushroom	11. Soda-pop can (book)	Soda-pop can has lemonade inside
12. Curly's nut (book)	Nut moved from tree to bush	12. Fish food box (book)	Fish food box has marbles inside

[a]All Locations tasks involved two characters, two locations, and an object that was transferred between them. For example, "Marcia's muffins" used two dolls, a tray of muffins, a refrigerator and a table with drawer. Book tasks showed key events with colored illustrations. For example, "Billy's rabbit" showed (a) Billy playing with his rabbit, (b) Billy putting the rabbit in the rabbit house, (c) Billy leaving, (d) the rabbit emerging from the rabbit house and hopping under the bed, (e) Billy returning to check on his rabbit, (f) Billy centered between the two locations ("Where will he look?"), and (g) Billy standing next to the rabbit house ("He's looking in the rabbit house"). [b]All Contents tasks involved two characters (one who looks inside the container and one who is naive) and a familiar container holding noncanonical contents. For example, the "Playdoh can" prop task used two dinosaur figurines and a playdoh can containing a rubber ball. Book tasks showed the key events with illustrations. For example, "Celine's lunchbox" showed (a) Celine, described as "hungry for lunch" with hand on stomach, (b) Celine holding her lunchbox, (c) Celine opening lunchbox to find a doll inside, (d) Celine with lunchbox closed up again, (e) Celine's friend Jamie (naive character) entering, (f) Jamie regarding the lunchbox ("What will he think is inside here?), and (g) Jamie pointing to the lunchbox, saying "Hey, I bet there's some yummy lunch in there."

tents of a crayon box that actually contained keys. To pass these tasks, children had to respond correctly to both the target question and the reality control question (e.g., "Where is the penny really?"; "What is really in this box?"). In the Seeing–Knowing task (similar to Pratt & Bryant, 1990), children were shown two dolls and a closed plastic can. The researcher acted out one doll opening and looking into the can, and the other picking up (but not opening) the can. Children were asked "Which one knows what's inside?" and, as a memory control, "Which one looked in the can?" Children had to be correct on both questions to pass. In the FB Explanation task (similar to Bartsch & Wellman, 1989), a toy giraffe ("Geoffrey"), who wanted bandaids, was shown searching for them in an empty bandaid box, while an unmarked box full of bandaids stood nearby. Children were asked, "Why is Geoffrey looking for the bandaids *here*?"; "What does he think?"; and as a control, "Where are the bandaids really?" We coded children's responses into categories based on whether they included appropriate references to the character's underlying beliefs/knowledge, or whether they referred to other aspects of the story, such as the character's desires or the bandaids' real location (see Results for final coding categories).

Inclusion criteria. Only children performing below chance (< 2 correct) on the three standard FB tasks could participate. Of 36 children included in the study, 25 (69%) failed all three standard tasks and the rest failed two out of three. All children in the study showed good performance on the control questions, with no child failing more than one.

Microgenetic sessions. Microgenetic sessions began within 1 to 2 weeks of a given child's pretest. Children in the *microgenetic* group participated in a total of 12 microgenetic sessions, with sessions occurring twice a week over a 6- to 7-week period and consecutive sessions separated by at least 1 day. Due to absences and school closings, sometimes children had longer lags between sessions than planned, but all children received a comparable experience of 12 sessions spaced over about 6 to 7 weeks of time.

During each microgenetic session, children received two FB tasks (one Locations, one Contents) for a total of 24 FB tasks. Tasks always focused on other people's (not the child's own) FBs. Different settings and central characters were used for each task. Half were presented as storybooks and half were acted out with props. Presentation order was counterbalanced across sessions for task type and format.

For each task, the researcher presented the FB scenario, asked the child to predict the outcome, and then provided "implicit feedback" by revealing the actual outcome of the story. In Locations tasks, the researcher first asked where the character would look for the desired object (e.g., "Where will Marcia look for her muffins?"). Following the child's response she said, "Well, let's see where he/she is going to look," and then revealed the character searching in the original location. In Contents tasks, the researcher asked what the character would think was inside the container, and following the child's response said, "Well, let's see what he/she says," and then had the character report the canonical contents of the container (e.g., "He says, 'Hey, here's some playdoh!'"). We refer to this feedback as "implicit," because children were not told explicitly if their responses were right or wrong.

Next, a *reality control question* (e.g., "Where are Marcia's muffins really?") served to confirm that children were following the stories and helped segue into the explanation request. The researcher then called the child's attention to the mismatch between reality and the character's actions or thoughts (searching in the wrong location, stating the wrong contents of the container) and asked the child to explain this event—for example, "So Marcia's muffins are really in the *refrigerator*. But Marcia is looking for them in the *drawer*! Why is she looking there?" If children responded by appropriately referring to the character's beliefs or knowledge (e.g., "She doesn't know where her muffins are," "She thinks they are in there"), they were not questioned further. However, if children responded in some other way (e.g., by describing details of the current situation, by referring to the character's desires, or by saying "I don't know") they were asked for an explanation a second time (e.g., "Yes, Marcia wants her muffins. But she's looking for them way over here! What happened?"). If children did not respond with an appropriate response to the second request, they received the final question, "What does he/she think?"[1]

So that children did not hear the same exact question repeated over and over again, several different wordings were used across the tasks. Sometimes the initial explanation question was a specific "why" question, as in the aforementioned example, and other times it was a more general "what happened" question that referenced the character's wrong or mistaken thoughts/actions (e.g., "Is she looking in the wrong place? What happened?" or "Did she make a mistake? How did that happen?"). These prompts were counterbalanced across the tasks so that all children received them in a standard way. Within a given session, for example, children received one specific "why" question and one "what happened" question, with the order of these prompts alternating across tasks. We hoped that including multiple kinds of explanation requests would keep children engaged in the task of explanation-seeking over the microgenetic sessions.

At the end of each task, children were asked one or two brief memory questions. We did this to include simple questions about the tasks that children could answer easily. Such questions asked, for example, whether the focal object had moved during the story (Locations), where it was first and where it went next (Locations),

[1]For Contents FB tasks, explanation requests focused on the discrepancy between what was really in the container and what the character reported was in there (e.g., "So there's really a ball in this playdoh can. But Sammy said it's *playdoh* in here. Why did he say that?"). In both cases, children are asked to explain behavior derived from a FB, rather than to explain the source of the FB itself.

whether the character had ever seen inside the container (Contents), or what was usually found inside such containers (Contents).

In addition to FB tasks, children received one true belief (TB) task every other session for a total of six TB tasks, with order of presentation counter balanced across sessions. TB tasks were nearly identical to FB tasks; however at no point in these stories did a character possess a FB. Thus, in Contents tasks a familiar container was opened to reveal appropriate contents, and in Locations tasks the object was removed and then replaced in its original location. TB tasks were included so that children would not always expect our tasks to involve a trick. They also allowed us to track children's performance on a task with a similar story structure, but no FB content. Children were not asked to explain outcomes in TB tasks.

Like the microgenetic group, the *comparison* group also received a series of microgenetic sessions over the course of several consecutive weeks, using the same kinds of FB and TB tasks used with the microgenetic group. Like the microgenetic group, children in the comparison group received a total of 24 FB tasks, with two microgenetic sessions per week and consecutive sessions separated by at least 1 day. They also received implicit feedback about the accuracy of their predictions for each FB task, just as in the microgenetic group.

However, there were two important differences in the experiences the two groups received. First, the tasks were spaced differently. Whereas the microgenetic group received the 24 tasks as 12 sessions of 2 FB tasks (one Contents and one Locations) each, and the comparison group received them as 6 sessions of 4 FB tasks (two Contents and two Locations) each. Thus, although both groups of children received the same total number of tasks, the microgenetic group's schedule of FB experiences was more distributed over time.

Second, the requests for FB explanations differed across the two conditions. Children in the microgenetic group were asked for explanations on every FB task, but those in the comparison group were asked for explanations on only *half* of the tasks (i.e., one Contents and one Locations task per session). Similar to the microgenetic group, the comparison group received a basic explanation request followed by a second request and a "think" question if needed. However, the wording of the explanation requests did not vary; a basic "why" request was repeated twice before the final question, "What does he/she think?"[2]

---

[2]These differences in microgenetic sessions across the two conditions reflected an early hypothesis that children in the comparison group were not improving on the FB tasks because they were growing bored or inattentive during the sessions. To forestall this problem in the microgenetic condition, we reduced the number of tasks per session by one half (while keeping the number of explanations per session the same) and introduced a greater variety of explanation requests. Subsequent analyses to assess children's affective engagement during testing, however, did not find significant group differences (see Results), which suggests that boredom or disengagement was not a primary reason for the comparison group's lack of improvement.

The *control group* did not participate in any testing between pre- and posttest. the number of intervening weeks between pre- and posttest for the control group was matched to that of the microgenetic group, which was the longer of the two microgenetic conditions.

Posttest. Children in the two microgenetic conditions received posttests roughly 1 week following their last microgenetic session (*M* = 9.0 days, *SD* = 4.0). The comparison group averaged 6.5 weeks (*SD* = 0.8) between pre- and posttest and the microgenetic group averaged 9.0 weeks (*SD* = 1.5). The timing of the posttest for the control group (8.9 weeks; *SD* = 1.2) was matched to that of the microgenetic group.

The posttest consisted of parallel versions of the three standard FB tasks used on the pretest. In addition, children were retested on the FB Explanation and Seeing–Knowing tasks. There were also two novel tasks used to see whether children might generalize new FB understandings to related concepts. The first was an Appearance–Reality task like those used by Flavell et al. (1986). Children were shown two different objects: a sponge that looked like a rock and a white toy egg that appeared blue under transparent blue plastic. For each object, children were asked how the object looked "when you look at it right now" and what the object was "really and truly." Appearance–Reality tasks show a similar developmental trajectory as FB tasks and the two are correlated in prior research (e.g., Gopnik & Astington, 1988). The second task tested children's understanding of the link between FB and emotion (adapted from Experiment 1 of Harris, Johnson, Hutton, Andrews, & Cooke, 1989). Props were used to act out a story in which George, a tricky monkey, gets a bag of his friend Ernie's favorite candy, replaces the contents with rocks, and then leaves it on Ernie's table. Children were asked: (a) "How will Ernie feel when he first sees this bag, before he looks inside it—happy or sad?" (b) "What is really inside this bag?" and (c) "How will Ernie feel when he opens the bag and finds out there are rocks inside–happy or sad?" This task may be somewhat harder than standard FB tasks; Harris et al. (1989) reported that children do not show consistent success on it until age six.

RESULTS

Data were analyzed at both the group and individual levels to document the overall effects of the experimental conditions as well as examine change processes in detail.

Group-Level Analyses

Improvements from pretest to posttest. Figure 1 shows the three groups' performance on the four pass–fail tasks used at pretest and posttest (Locations,

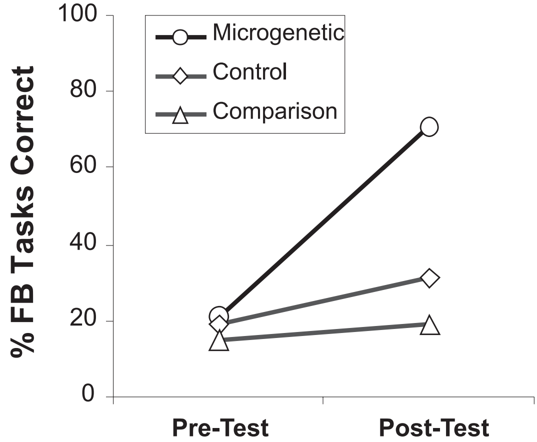

FIGURE 1 Pretest and posttest performance: Average pre- and posttest performance of microgenetic, control, and comparison groups, reported as percentage of false belief tasks correct out of four total tasks.

Contents–Other, Contents–Self, and Seeing–Knowing). At pretest, all three groups had equally low levels of FB task performance (21%, 19%, and 15% correct for microgenetic, control, and comparison groups, respectively), $F(2, 33) = .381$, *ns*.[3] At posttest, however, the groups differed, $F(2, 33) = 19.38$, $p < .001$. The microgenetic group (71% correct) outperformed both the control group (31% correct), $t(22) = 4.29$, $p < .001$, and the comparison group (19% correct), $t(22) = 6.42$, $p < .001$. The control and comparison groups did not differ, $t(22) = 1.41$, *ns*. Within-subject analyses confirmed that significant improvement in performance occurred only for the microgenetic group, $t(11) = 4.69$, $p < .01$, and not for the other two groups.

Posttest improvements by task type. Improvements were more pronounced for certain FB tasks, as shown in Table 2. In the microgenetic group, the greatest gains from pre- to posttest occurred for task types used in the microgenetic sessions. The microgenetic group improved from 17% to 75% correct on the Locations task, McNemar's $\chi^2(1, N = 12)$, $p < .05$, and from 8% to 83% correct on the Contents–Other task, McNemar's $\chi^2(1, N = 12)$, $p < .01$. Notably, the control and comparison groups did not improve on either of these tasks.

[3]Analyses were run on the total number of tasks passed (ranging from 0–4), but for ease of interpretation we present results in terms of overall percentage correct. For all analyses, "*ns*" denotes $p > .05$.

TABLE 2
Proportion Correct on Pretest and Posttest Tasks by Group

	Microgenetic		*Comparison*		*Control*	
	Pretest	*Posttest*	*Pretest*	*Posttest*	*Pretest*	*Posttest*
Locations	.17	.75*	.00	.17	.33	.33
Contents–Other	.08	.83**	.08	.00	.17	.25
Contents–Self	.00	.58*	.08	.00	.00	.00
False Belief Explanation[a]	–.33	.50*	–.58	–.42	–.42	–.08
Seeing–Knowing	.58	.67	.42	.58	.25	.67
False Belief Emotion	—	.17	—	.00	—	.08
Appearance–Reality	—	.75	—	.58	—	.75

[a]Mean level of sophistication, rather than proportion correct.
*$p < .05$. **$p < .01$.

TABLE 3
Types of False Belief Explanations

Explanation	*Examples*
Belief	"He doesn't know carrots are in there."
	"He thinks it's under there, but it's not."
Mistake	"She made a mistake."
Desire	"She loves soda pop."
	"She wants to feed her dog."
Situational	"There isn't any playdoh in there."
	"His friend took it."
Don't know (or no response)	"I don't know."

On the FB Explanation task, children's explanations were coded into five categories, referred to as Belief, Mistake, Desire, Situational, and Don't Know explanations (see Table 3). Explanations that described the protagonist's beliefs or knowledge were considered the most sophisticated, followed by those referring to a mistake, by those describing his desires, and then by those that mentioned aspects of the situation. When children gave more than one kind of explanation on the task, we used the highest level response in analyses. Explanations were coded by two independent raters; interrater agreement was 98%.

At pretest, Situational explanations were the most frequent response type and were equally common in all three groups, given by 7 of 12 children (58%) in the microgenetic group, 8 of 12 children (67%) in the control group, and 9 of 12 children (75%) in the comparison group, $\chi^2(2, N = 36) = .75$, *ns*. Belief explanations were the least frequent, given by 3 children (25%) in the microgenetic group, 3 children (25%) in the control group, and 2 children (17%) in the comparison group. By posttest, however, situational explanations were lower in the microgenetic group than in the other two groups, with only 1 child (8%) in the microgenetic group giving a Situational explanation, and 5 (42%) in the control group and 7 (58%) in the comparison group doing so, $\chi^2(2, N = 36) = 6.74, p < .05$. At posttest, the most common response type in the microgenetic group was Belief explanation, with 6 children (50%) giving Belief explanations and 1 additional child giving a Mistake explanation. Comparing higher level (Belief/Mistake) explanations across groups at posttest, the microgenetic group (58%) gave more Belief/Mistake explanations than the comparison group (17%), $\chi^2(2, N = 36) = 4.44$, $p < .05$, but did not differ significantly from the control group (33%), $\chi^2(2, N = 36) = 1.51$, *ns*.

These initial findings suggest that improvement in explanations from pre- to posttest involved both decreased Situational explanations and increased Belief/Mistaken explanations. To create a combined measure that reflected both children's movement away from incorrect Situational explanations and their movement toward correct mentalistic explanations, we gave each child an explanation score for pre- and posttest as follows: –1 for Situational explanations, 0 for Don't Know or Desire explanations, and +1 for Belief/Mistake explanations. Means for the three groups are given in Table 2. As the table shows, on this combined measure, only children in the microgenetic group changed significantly from pre- to posttest, $t(11) = 2.80, p < .05$.

The microgenetic group also showed clear improvement on a FB task *not* used during microgenetic testing: the Contents–Self task, which asked children to report their own prior FB. The microgenetic group improved from 0% to 58% correct on this task from pre- to posttest, McNemar's $\chi^2(1, N = 12), p < .05$. The other groups did not improve. In fact, no child in the control or comparison groups passed this task at posttest.

The remaining posttest tasks did not show group differences in performance. For the Seeing–Knowing task, children in the microgenetic group did not outperform other groups, $\chi^2(2, N = 12)$, = .24, *ns,* nor did any of the groups show within-subjects improvement, McNemar's $\chi^2(1, N = 12)$, all *p*s > .05. Similarly, all three groups performed equally well on the Appearance–Reality task, $\chi^2(2, N = 12) = .14$, *ns.* There were also no group differences in performance on the posttest FB Emotion task, $\chi^2(2, N = 12) = 2.18$, *ns.* We had anticipated that some children who improved on standard FB tasks might also improve on these other tasks. However, it may be that these particular tasks were not at the proper level of difficulty to show transfer effects. For example, in all groups, performance on the Seeing–Knowing and Appearance–Reality tasks was actually quite good (see Table 2) relative to children's very low performance on accompanying standard FB tasks. At the opposite extreme, very few children in any group passed the FB Emotion task. This task may have been too difficult even for children with an initial understanding of FB (see Harris et al., 1989).

Performance improvements during microgenetic testing. Figure 2 displays the performance of the microgenetic and comparison groups over the microgenetic sessions. Recall that both groups received the same total number of FB tasks (24), despite the fact that tasks were spaced differently over time for each group. To compare performance across groups based on equal amounts of task experience, Figure 2 reports the data in terms of consecutive "task blocks," each consisting of four FB tasks (two Contents and two Locations). For the comparison group, these task blocks are identical with their microgenetic sessions. For the microgenetic group, each task block shows performance over two consecutive sessions.

Children in the microgenetic group improved significantly from their first microgenetic task block (14% correct) to their last (67% correct), $t(11) = 3.57$, $p < .01$. In contrast, children in the comparison group did not improve over the course of testing, $t(11) = 0.56$, *ns.* A factorial analysis of variance (ANOVA) using the six task blocks to create a within-subjects factor (time) confirmed significant main effects for condition (microgenetic or comparison), $F(1, 22) = 44.52$, and for time, $F(5, 110) = 8.75$ and a significant Time × Condition interaction, $F(5, 110) = 9.33$, all ps < .001. On average, children in the microgenetic condition showed FB task improvement by the second task block (compare their 40% correct performance to the comparison group's 10%) and reached above-chance performance (71% correct) by block four.

Despite differences in performance on FB tasks, children in both groups showed high performance on true belief tasks throughout microgenetic testing; the microgenetic group was correct on 93% of their true belief tasks, the comparison group on 96%. Similarly, children showed very high levels of performance on real-

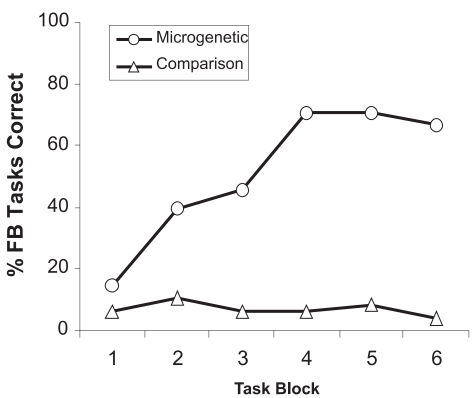

FIGURE 2 Performance during microgenetic sessions: Average performance of microgenetic and comparison groups during microgenetic sessions, reported as percentage false belief tasks correct out of four total tasks per task block.

ity control questions for FB tasks; the microgenetic group was correct on 95% of the control questions, the comparison group on 91%.

FB explanation during microgenetic testing. During their microgenetic sessions, children were asked to explain FB events with a series of two general explanation requests and a final "think" question directly asking about the character's thoughts. For responses to general requests, children's explanations were coded into one of five basic categories: Belief, Mistake, Desire, Situational, and Don't Know explanations (see Table 3). Responses to direct questions about the character's thoughts were scored as correctly mentalistic if they were consistent with the character's thoughts or knowledge in the situation (e.g., "He thinks his bunny's in there" or "He doesn't know where it is" when a character searches in the original location), and incorrect if they were not (e.g., stating that the character thinks the object is in the location not being searched, describing his desire for the object, saying "I don't know," etc.). One person coded all of the children's responses, and then half were recoded by an independent rater as a reliability check. Interrater agreement was 98%.

We first assessed basic changes in FB explanations with a series of 2 × 2 (Condition: Microgenetic or Comparison × Time: First or Last Task Block) ANOVAs on the frequency of each explanation type: Situational, Don't Know, Desire, and Belief/Mistake. These yielded only two significant effects. There was a significant decline in Situational explanations over time in both groups (from 63% to 29% in the microgenetic group and from 58% to 37% in the comparison group), $F(1, 22) = 12.55$, $p < .01$, and the microgenetic group gave more Belief/Mistake explanations on average than the comparison group (19% vs. 2%), $F(1, 22) = 5.77$, $p < .05$. For performance on "think" questions, the microgenetic group showed better performance (57% correct) than the comparison group (17% correct) by the end of the first task block, $t(22) = 3.41$, $p < .01$, with no further changes in either group by the last task block (56% vs. 12% correct), ps > .05.

These basic analyses do not fully capture the differential experiences and differential success with FB explanation that children in the two groups accumulated over their microgenetic sessions. In addition to reflecting ongoing improvement in FB understanding, the microgenetic group's success in generating appropriate mentalistic explanations (both spontaneously and in response to direct questions about characters' thoughts) also defines an important difference between the microgenetic experiences of the two groups.

Figure 3 illustrates these differences in the groups' experiences over time. Consider first the panel for mentalistic explanations. Rather than frequency over time, this graph presents the *cumulative* number of mentalistic explanations that children in each group provided—that is, the number graphed for each task block sums all the mentalistic explanations each group generated up to that point. Mentalistic explanations for this tally include Belief/Mistake explanations generated sponta-

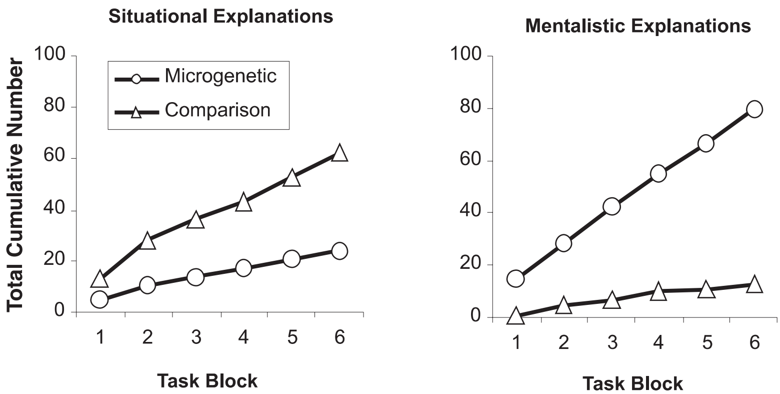

FIGURE 3 Changes in false belief (FB) explanations during microgenetic sessions: Cumulative number of situational and mentalistic FB explanations generated in microgenetic and comparison groups over time, adjusted for total explanation requests received.

neously as well as correct mentalistic responses to the final "think" question. Figure 3 shows that, beginning with small differences evident by the end of the first task block, by the end of all microgenetic sessions (i.e., by task block six) the cumulative experience with mentalistic explanation generated by the two groups is quite different.

Of course, the two groups did not have the same opportunities to make FB explanations. Microgenetic group children were asked to explain each and every FB task they received (a total of 24 separate explanation requests), whereas comparison group children were asked to explain every other FB task (a total of 12 requests). In Figure 3, the line for the microgenetic group has been adjusted for this difference by halving the numbers for each task block. Thus the (adjusted) line shows the explanation experiences of the microgenetic group, *as if* they had only 12 explanation requests like the comparison group. Even with this adjustment, it is clear that the cumulative experience of the two groups is very different; by the end of the third task block, microgenetic children had provided 42 mentalistic explanations whereas comparison children had provided only 7. An unadjusted line, were it depicted in Figure 3, would be doubly steep; by the end of the sixth task block, microgenetic children had in fact provided 160 mentalistic explanations but the comparison children had provided only 13.

The data for Situational explanations provides an important contrast, and in conjunction with the data for mentalistic explanations, reveal a clear divergence in the accumulating experiences of these two groups. As our initial analyses showed, although the session-by-session frequencies of Situational explanations declined over task blocks in both groups, Situational explanations did not disappear completely and so also accumulated over time. As shown in the left panel of Figure 3, by the end of the sixth task block, comparison children had generated more Situational explanations than microgenetic group children had.

One further analysis considered whether there was any relation between the use of specific explanation requests (i.e., "why" questions vs. "what happened" questions) and the explanations children generated in response. Collapsing across all explanation data for children in the microgenetic group, who received both kinds of requests, chi-square results showed that proportions of explanation types were the same for both cases. Specifically, Belief/Mistake explanations occurred equally in response to both "why" and "what happened" requests (15% and 11% of responses, respectively), as did Situational explanations (32% and 46%), Don't Know explanations (47% and 40%), and Desire explanations (6% and 3%), all ps > .05. Similarly, the use of the term "mistake" in some explanation requests did not account for children's use of such terms in their own explanations; Belief/Mistake explanations were no more likely in response to "mistake" explanation requests than to requests that did not include that term (9% vs. 14% of responses, respectively).

Engagement during microgenetic testing. Children's level of engagement during microgenetic testing was also assessed, as successfully learning from implicit feedback or FB explanation might depend on children's interest in or attention to the tasks. Engagement coding was conducted post hoc from video tapes of the sessions, and children received a 0 to 3 score for each of the 24 FB tasks they received. Three microgenetic group children and one comparison group child lacked video for two tasks, so their analyses were based on 22 rather than 24 tasks. All other children had complete data.

Signs of *engagement* included: full attention to the task or researcher, generally positive affect (e.g., smiling, laughing), and responding to questions readily and with enthusiasm. Signs of *disengagement* included: little attention to the task or researcher (e.g., looking/turning away); a distinct lack of positive affect or clear negative affect (e.g., frowning, verbally expressing dislike for tasks); and minimal participation in or active disruption of tasks (e.g., not responding to questions, leaving the testing table). For each FB task, children were scored as highly engaged (score = 3) if they clearly had all of the signs of engagement identified earlier; as mostly engaged (score = 2) if they had most of the signs of engagement, but did not appear maximally engaged for one reason or another; as somewhat engaged (score = 1) if they appeared engaged for the most part, but also showed signs of disengagement; and as disengaged (score = 0) if they showed more signs of disengagement than engagement. A rater naive to the study hypotheses and results performed the initial coding and a second rater recoded 15% of the data to establish reliability. Interrater agreement was 89%.

Children's individual engagement scores, averaged over tasks, ranged from a low of 1.7 to a high of 2.9. There were no differences in average engagement between the microgenetic ($M = 2.4$, $SD = .3$) and comparison groups ($M = 2.4$, SD = 0.3), $t(22) = .06$, *ns*, with engagement consistently high in both groups. A 2 × 2 (Condition × Time) ANOVA found small but significant decreases in overall engagement from the first task block to the last across both groups (*M*s = 2.6 vs. 2.2), $F(1, 20) = 6.24$, $p < .05$, but no condition or interaction effects.

Individual-Level Analyses

In sum, group-level analyses found that children in the microgenetic group improved on FB tasks and gained experience giving mentalistic FB explanations over time, whereas children in the comparison group did not. We now turn to more detailed analyses of individual patterns of change, with an eye to what these data reveal about the nature of children's successful or unsuccessful transition to FB understanding.

Patterns of improvement during microgenetic sessions—FB task performance. As in the group analyses, we tracked changes in children's FB task performance over the microgenetic sessions by plotting the number of tasks children passed in each of their six microgenetic task blocks. Figure 4 presents data for all children in the microgenetic and comparison groups.

We identified as "improvers" those children who achieved at least one task block of above-chance (75 or 100% correct) performance on the tasks. Of the 12 microgenetic group children, 9 met the improvement criterion, with 7 actually showing above chance or 100% correct performance on three or more consecutive blocks. Three microgenetic group children never achieved above-chance performance, and so were classified as "nonimprovers." All 12 comparison group children were below chance at all time points and so were also nonimprovers. We validated our improvement criterion by examining its relation to posttest performance. Collapsing across comparison and microgenetic groups, whether or not a child improved during the microgenetic sessions (where 0 = no improvement and 1 = improvement) was highly correlated with overall performance on the posttest, Spearman's $\rho = .79$, $p < .001$.

Because it was possible that some children might have improved on only one type of task, we also assessed improvement by task type. We considered a child to have improved on a given task if he or she passed that kind of task three or more times in a row during microgenetic testing. Prior microgenetic work has used similar criteria to identify children who show task improvement over time (e.g., Siegler & Svetina, 2002). Of the nine improvers in the microgenetic group, seven improved on both tasks, one improved only on Contents, and one improved only on Locations. The three nonimprovers in the microgenetic group and all 12 nonimproving children in the comparison group did not improve on either task.

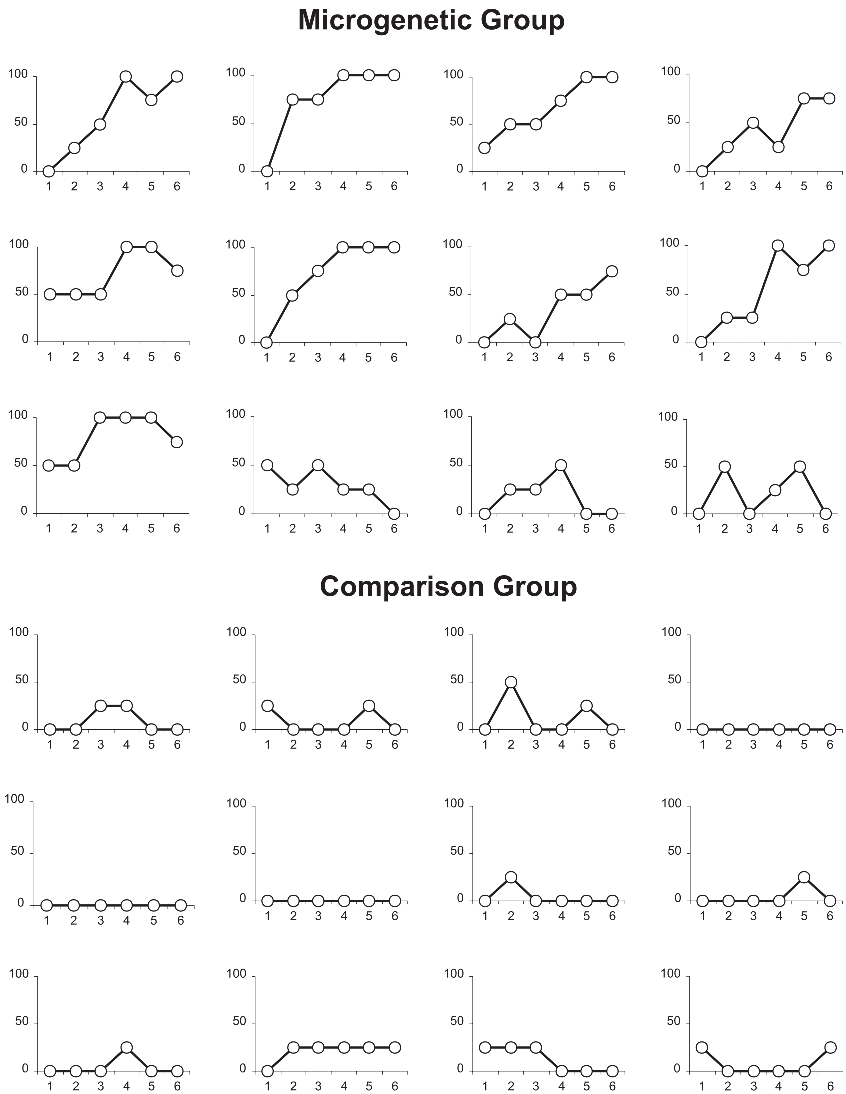

FIGURE 4 Individual children's false belief (FB) task performance during microgenetic sessions, reported as percentage of FB tasks correct out of four total tasks per task block.

These data confirm earlier group-level results showing improvement only for children in the microgenetic group. They also reveal intriguing patterns in how improvement, when it occurred, unfolded over time. We highlight three features of particular interest.

First, the data reveal that children's FB task improvement was gradual (see Figure 4), both in the sense that children required multiple task sessions before showing consistent improvement, and in the sense that children had lags between passing a task for the first time and showing consistent success (i.e., three passes in a row) on it. For example, improvers showed significant lags of 2.1 sessions (SD = 1.8) between their first pass and their first block of consistent success on a given task, demonstrating that children did not master these tasks quickly; one-sample $t(8) = 3.37$, $p < .05$.

Second, as these lags suggest, all improving children also showed regressions in performance where they failed tasks they had previously passed. The majority of these occurred prior to children's achieving consistent success on tasks—out of 19 total cases of regression following task success, 14 (74%) occurred prior to achieving consistent above-chance performance, whereas 5 (26%) occurred after it, $\chi^2(1, N = 19) = 4.26$, $p < .05$. That fluctuations were especially prevalent during the lag before children's mastery of the tasks provides evidence of a transitional period in children's theory of mind development, akin to that implied by the earlier meta-analysis (Wellman et al., 2001). As the individual microgenetic graphs indicate, no child advanced directly from 0% correct to 100% correct performance on the tasks; all proceeded through some period of intermediate success.

Finally, with respect to performance on the two different FB task types, children had similar trajectories of improvement on the two tasks, but did not achieve success on them simultaneously. Improving children had average lags of 2.8 sessions (SD = 1.7) between their first success on one task and their first success on the other, one-sample $t(8) = 4.86$, $p < .01$. All also exhibited a pattern of first attaining a level of consistent success on one task before attaining consistent success on the other. On average, improvers achieved success (i.e., three passes in a row) on one FB task type by about their fifth session (M = 5.3, SD = 1.9), but did not achieve consistent success on the second task type until about their ninth session (M = 9.4, SD = 2.2), $t(6) = 7.48$, $p < .001$. Children thus appeared to master FB concepts in one context before mastering them in another. We also note that improving children did not consistently pass one type of task first; six children passed Locations tasks before Contents, and three did the reverse.

Patterns of improvement during microgenetic sessions—FB explanations. We tracked changes in children's FB explanations using the coding categories described earlier. Children in both conditions typically generated multiple explanation types over the course of their sessions. Excluding Don't Know explanations, microgenetic group children averaged 2.4 (*SD* = 0.9) and comparison group children averaged 1.8 (*SD* = 0.9) different FB explanation types during microgenetic testing, $t(22) = 1.56$, *ns.* Of special interest was children's use of mentalistic explanations that referred to characters' beliefs, knowledge, or mistaken thoughts and actions.

We first asked whether improvement in FB task performance during microgenetic sessions was associated with giving mentalistic explanations, either in the form of children's "spontaneous" explanations following general explanation requests ("Why is he looking there?"), or in the form of consistently correct responses on direct questions about characters' beliefs ("What does he think?"). In terms of individual children, there were significantly more children in the microgenetic group who ever gave mentalistic FB explanations to general explanation requests or performed above chance on direct questions about characters' beliefs than there were in the comparison group (10 vs. 4), $\chi^2(1, N = 24) = 6.17, p < .05$. There was also a strong relation between FB task improvement and mentalistic explanation, $\chi^2(1, N = 24) = 6.99, p < .01$. That is, eight of nine improving children in the microgenetic group made references to characters' mistakes, beliefs, or knowledge following a general explanation request—on average, giving such responses on 28% of tasks they were asked to explain. Only 5 of 15 nonimproving children in either group gave such responses, on average making them on 13% of tasks they were asked to explain.

To assess individual patterns of change in FB explanation, we also identified for each child the explanation types that showed the largest increase and largest decrease in frequency from the first to last task block. Children's patterns fell into four main categories of change. Two were patterns of nonimprovement—children either showed no change in their FB explanations over time or showed an increase in Situational explanations. Two were patterns of (either clear-cut or more modest) improvement in FB explanation. In the pattern of clear improvement, children had decreases in lower level explanations (e.g., Situational, Don't Know responses) and increases in higher level, mentalistic explanations. In the pattern of less straightforward improvement, children showed decreases in Situational explanations with increases in Don't Know explanations.

All nine improvers in the microgenetic group showed improvements in their FB explanations over time. Five showed a clear pattern of improvement toward more mentalistic explanations, and four showed decreasing Situational explanations and increasing Don't Know explanations. The three nonimprovers in the microgenetic group either gave more Situational explanations over time or showed no change. In the comparison group, no child showed clear improvements in either FB task performance or FB explanations; however, four children did show decreasing Situational explanations and increasing Don't Know explanations over time. These results suggest that improvement in FB task performance and FB explanation were related in the sense that FB task improvement was always accompanied by changes in FB explanation, although the reverse was not always true.

At a more micro level, we also considered how FB explanation and prediction unfolded over time. Again, we focused on patterns of change in the nine improving children from the microgenetic group, looking at performance on FB tasks, general explanation requests, and direct questions about characters' thoughts. Improving children showed a clear pattern of first achieving success on direct questions about characters' thoughts *before* providing their own mentalistic explanations to general requests. On average, improvers correctly answered direct questions about what characters were thinking by their first or second session ($M = 1.2$, $SD = 0.4$), but did not spontaneously offer mentalistic explanations of their own until the second or third session ($M = 2.7$, $SD = 1.4$), $t(7) = 3.27$, $p < .05$, nor did they pass FB tasks until the second or third session ($M = 2.9$, $SD = 1.3$), $t(8) = 3.78$, $p < .01$.

Descriptively, six of nine improvers gave their first spontaneous mentalistic explanations prior to passing any FB tasks, two children showed the reverse, and one passed tasks but did not give mentalistic explanations. However, mentalistic explanations always preceded achieving at least one pass on both kinds of tasks ($M = 6.1$, $SD = 2.0$), $t(7) = 3.83$, $p < .01$. For nonimprovers, only those children who had early success on direct questions about characters' thoughts ever gave their own mentalistic FB explanations, although these children did not proceed to FB task success as improvers did. Thus, in our data, success on direct questions about characters' thoughts preceded spontaneous mentalistic explanation, which also preceded consistent success in FB prediction. These results are generally consistent with earlier cross-sectional reports that 3-year-olds with "transitional" understandings of FB can often explain FB events before they are able to predict them (Bartsch & Wellman, 1989).

Predicting improvement in FB understanding. We used regression modeling to further test relations between improvement outcomes and aspects of children's performance during the microgenetic sessions. We expected that condition effects would largely account for the observed outcome differences. However, person variables such as age, gender, and pretest performance, and aspects of children's behavior during microgenetic testing, such as the type and variety of FB explanations given, may also have played a role. We considered two main outcome variables: *Posttest performance* (total performance on Locations, Contents–Self, Contents–Other, and Seeing–Knowing tasks at posttest, scored 0–4), and *FB task improvement* (whether children improved on neither, one, or both Locations and Contents FB tasks by the end of their microgenetic sessions, scored 0–2).

Preliminary analyses examined associations among the following predictors: condition (microgenetic or comparison), child's age in months, child's gender, two pretest measures (number of tasks passed, whether child gave a mentalistic response on the FB Explanation task), average level of engagement during microgenetic sessions, and three variables describing children's explanations during the first three microgenetic task blocks (number of different FB explanation types given, whether child was above chance on direct "think" questions, and number of mentalistic FB explanations given). As Table 4 shows, there were no significant correlations between age, gender, pretest performance, and engagement and any of the other predictors or outcome measures. As expected, being in the microgenetic group was associated with FB prediction improvement and better posttest performance, as well as with mentalistic explanation and better performance on "think" questions during initial microgenetic sessions. In turn, explanation variables were strongly correlated with performance on outcome measures.

Stepwise linear regressions identified the predictors most reliably associated with improvement. Results are given in Table 5. For posttest performance, condition and mentalistic explanation accounted for 71% of the variance, $F(2, 21) = 28.4$, $p < .001$. Condition alone accounted for 64% of the variance, $F(1, 22) = 41.2$, $p < .001$, and mentalistic explanation significantly increased the model fit to R^2 (adjusted) = .71, $F_{\text{inc}}(1, 21) = 6.11$, $p < .05$. These factors also accounted for 67% of the variance in FB task improvement, $F(2, 21) = 28.3$, $p < .001$. Condition alone accounted for 53% of the variance, $F(1, 22) = 27.1$, $p < .001$, and mentalistic explanation significantly increased the model fit to R^2(adjusted) = .67, $F_{\text{inc}}(1, 21) = 10.2$, $p < .01$. In both cases, children were more likely to perform well if they were in the microgenetic condition and if they gave more spontaneous mentalistic explanations during microgenetic sessions. None of the other variables significantly predicted improvement.

These relations between predictors and improvement outcomes were not reducible to preexisting differences in children's FB knowledge or explanation abilities. To reiterate, the microgenetic and comparison groups did not differ on the key variables of age, pretest performance, or engagement, and there were no significant correlations between age, gender, pretest performance, or engagement and other predictor variables, nor between them and improvement outcomes (see Table 4). Moreover, in the regression analyses, these variables did not predict to outcomes even when controlling for condition effects. (Incidentally, explanation performance in the *very first* microgenetic task block, as opposed to the first three blocks, also did not predict improvement, again suggesting that the crucial differences emerged only as sessions progressed.)

We also considered predictors of *FB explanation improvement* during microgenetic sessions, where children were given a 0 to 2 score denoting whether they showed no change in explanation, showed decreasing Situational explanations only, or showed both decreasing Situational explanations and increasing Be-

TABLE 4
Correlations Between Predictor Variables and Improvement Outcomes

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1. Posttest performance	—										
2. False belief task improvement	.80***	—									
3. False belief explanation improvement	.51*	.70**	—								
4. Age	.20	.04	–.19	—							
5. Gender (0 = female, 1 = male)	–.08	–.29	–.34	.07	—						
6. Pretest performance (4 tasks)	–.11	–.21	–.23	.01	–.16	—					
7. Pretest false belief explanation	.08	–.15	–.20	.04	.24	–.08	—				
8. Condition (microgenetic/comparison)	.81***	.74***	.52*	.01	–.09	.17	.10	—			
9. Mentalistic explanation	.62**	.68***	.66***	.19	–.31	–.25	–.17	.46*	—		
10. Performance on “think” questions	.60**	.44*	.24	.08	.07	.01	.35	.64**	.30	—	
11. Number of explanation types	.48*	.40	.52**	–.03	–.20	–.31	.04	.32	.62**	.21	—
12. Engagement	–.08	–.11	.17	.20	–.19	.38	.06	.01	–.15	.07	–.28

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 5
Results of Regression Analyses Testing Explanation as Mediator of the Relation Between Condition and Improvement

Dependent Variable	*Model*	*Variables in Model*	β	R^2 *(Adjusted)*	*Change in* R^2
Posttest performance	1	Condition	.807***	.64***	.08*
	2	Condition,	.661***	.71***	
		Mentalistic Explanation	.316*		
False belief task improvement in microgenetic sessions	1	Condition	.743***	.53***	.15**
	2	Condition,	.543**	.67***	
		Mentalistic Explanation	.432**		
False belief explanation improvement in microgenetic sessions	1	Condition	.517**	.23***	.14*
	2	Condition,	.393*	.35***	
		Number of Explanation Types	.395*		

*$p < .05$. **$p < .01$. ***$p < .001$.

lief/Mistake explanations. As shown at the bottom of Table 5, two factors accounted for 35% of the variance in FB explanation improvement, $F(2, 21) = 7.24$, $p < .01$. Condition alone accounted for 23% of the variance, $F(1, 22) = 8.03$, $p < .05$. The second factor was the total number of different explanation types children gave, which significantly increased the model fit to R^2(adjusted) = .35, $F_{inc}(1, 21) = 5.00$, $p < .05$. Children who showed more variety in their FB explanations were more likely to improve.

Finally, we considered the extent to which mentalistic explanation might mediate the effects of condition on improvement. According to Baron and Kenny (1986), establishing mediation first requires that condition be related to improvement outcomes. As shown in Table 5, this was true for both of our performance outcome variables: posttest performance, R^2 (adjusted) = .64, $F(1, 22) = 41.2$, $p < .001$; and FB task improvement, R^2(adjusted) = .53, $F_{inc}(1, 22) = 27.1$, $p < .001$. Second, condition must be related to mentalistic explanation. When condition was regressed on mentalistic explanation, it significantly predicted explanation patterns, R^2(adjusted) = .18, $F(1, 22) = 6.00$, $p < .05$. Third, the strength of the relation between condition and improvement outcomes should be reduced when mentalistic explanation is included in the model. As shown in Table 5, betas for condition were consistently reduced when mentalistic explanation was included. Nonetheless, in both cases the condition effect remained strongly significant in the second model, which is consistent with *partial* mediation. Thus, observed group differences in improvement during microgenetic testing are partly attributable to the different patterns of explanation that these conditions elicited. However, other differences between the conditions must also have affected children's performance.

DISCUSSION

Using a microgenetic approach, we successfully fostered and documented the development of FB understanding in a group of initially naive 3-year-old children. Children in our focal microgenetic group improved significantly over the course of their microgenetic sessions, achieving high posttest performance on both kinds of FB tasks used in the sessions (Locations and Contents–Other) and on an unfamiliar FB task (Contents–Self). They also showed development in their ability to explain FB events, as more children gave Belief/Mistake explanations and fewer gave Situational explanations over time. In contrast, children in the time-lag control group, who received no microgenetic experiences between pre- and posttest, showed no improvement on these measures over time. Children in the comparison group, which received a different set of microgenetic experiences, also did not improve.

Findings from two other microgenetic studies of theory of mind are consistent with key aspects of our results (Flynn et al., 2004; Wahl, 2001). Results from Flynn et al. (2004) mirror those from our control group, and results from Wahl (2001) mirror those from our comparison group. In Flynn et al.'s microgenetic study, 21 children initially aged 3;1 to 3;10 ($M = 3;5$) received multiple FB tasks designed to assess development in FB understanding without enriching children's experiences to promote more rapid theory of mind development. Children were tested six times, with tests spaced at 4-week intervals. Similar to our control group, who received only two widely spaced assessments, Flynn et al. did not observe significant improvements in children's performance on standard FB tasks over the course of their study.

Wahl (2001) hoped to promote theory of mind development with a microgenetic approach. Over a series of 10 weekly sessions, 36 children aged 3;0 to 3;7 ($M = 3;3$) received six different theory of mind tasks per session (mostly Locations and Contents). On initial tasks at each session, children received explicit performance feedback following their responses—the researcher either confirmed their answers or corrected errors. Following each feedback task, a similar task with no feedback assessed children's level of FB understanding. Wahl found that despite repeated exposure to the tasks (children received a total of 60 tasks over the course of testing), as well as explicit feedback on half the tasks, children did not improve over time. Similarly, children in our comparison group participated in repeated FB tasks where they received implicit feedback about correct responses, yet showed no progress in FB understanding. These results suggest that the everyday course of theory of mind development is not readily accelerated by exposing children to FB tasks. When no special attempts are made to foster theory of mind development, as in our control group and in Flynn et al. (2004), little change occurs and development unfolds over an extended period of time. Even exposing children to numerous FB tasks in the context of informative corrective feedback can be surprisingly ineffective in promoting change, as results from our comparison group and Wahl (2001) showed.

The somewhat larger literature investigating short-term training experiences also reports mixed success. Several studies show clear improvements in children's theory of mind task performance in at least some experimental conditions (e.g., Appleton & Reddy, 1996; Clements, Rustin, & McCallum, 2000; Lohmann & Tomasello, 2003; Slaughter, 1998). Others, despite reasonable training approaches, report no improvement or improvement restricted to trained tasks (Guajardo, 1999; Knoll & Charman, 2000). One difficulty in determining the extent of training gains in such studies is that they have not used rigorous pretesting to establish children's initial level of FB understanding. Training studies have typically screened children with only one FB task (or two trials of a single kind of FB task), leaving open the possibility that the short-term changes they report are occurring in children who already have somewhat developed knowledge. Children included in our study performed poorly on four different FB prediction and explanation tasks at pretest. Thus, we can be certain that their task improvements occurred despite very impoverished initial understandings.

As in microgenetic work, training conditions where children are simply exposed to relevant tasks do not tend to yield much improvement. Lohmann and Tomasello (2003) conducted the most comprehensive training study to date, comparing children in five different training conditions where children received a total of three training sessions. They observed no posttest improvement in FB understanding among children who received extensive experience with deceptive objects but had no discussions about them with the adult researcher. In contrast, their most effective training condition was one in which children had experience with the objects and the adult also used perspective-shifting discourse to highlight appearance–reality contrasts and model mental state sentence structures. Similarly, Clements et al. (2000) found that simply giving children corrective feedback about their performance on FB tasks did not lead to posttest improvement (but see Slaughter, 1998; Slaughter & Gopnik, 1996), but children who received conceptual explanations from the adult researcher about why their answers were correct or incorrect did improve. Appleton and Reddy (1996) and Knoll and Charman (2000) also reported success with discourse-based approaches where children were asked to explain aspects of FB scenarios and were then given additional conceptual information if their explanations were incomplete or incorrect. Furthermore, Appleton and Reddy (1996) reported an analysis relating children's posttest improvement to their ability to explain FB events during training. In their study, just as we document more extensively with our microgenetic analyses, children who gave more correct explanations for FB events during training were more likely to improve at posttest.

The Nature of Children's Improvement

Our microgenetic data show that children developed an understanding of FB gradually over the course of multiple weeks. Improvement was characterized by transitional periods between initial demonstrations of FB understanding and the achievement of consistent FB task success, as children progressed from first giving correct responses to direct questions about characters' thoughts, to producing their own spontaneous mentalistic explanations for FB events, to finally predicting FB task outcomes with increasing accuracy. In addition, all improving children had periods of fluctuating task performance where they failed tasks they had previously passed. These results demonstrate that children's understandings developed along a protracted path. New competencies did not emerge abruptly, nor did they completely replace previous, less-sophisticated modes of responding.

Several aspects of our data suggest that task improvements among children in the microgenetic group were genuine changes in children's understanding of FB, rather than reflecting lower level response strategies. First, if performance improvements were simply due to adoption of some simple response rule, such as a rule about always picking the location where the object was *not* currently located, we might expect to see *decrements* in true belief task performance as well as improvements in FB task performance when children were developing the rule. However, children were uniformly good on true belief tasks at all time points even as their FB task performance improved. Second, children who improved on FB tasks used during microgenetic sessions also extended what they learned to a different kind of FB task, one that required introspecting about one's own prior FB rather than predicting someone else's FB. Improvements on the Contents–Self task at posttest were apparent only among children who improved during microgenetic testing; other children did very poorly on the task. Out of 9 improving children, 7 (78%) improved on this task, but only 1 out of 15 (6%) nonimproving children did. Finally, children's responses to FB explanation requests provide evidence of their deeper understanding. Children whose FB task performance improved over time gave explanations for FB events that demonstrated an emerging awareness and understanding of mentalistic concepts underlying the task.

Accounting for Improvement

Our two microgenetic conditions contained many similar features, yet the results were strikingly different. Because our microgenetic data allowed us to examine change unfolding over time, we were able to examine how variation in children's experiences during microgenetic testing might serve to explain variation in patterns of change.

Results suggested that differences in the amount and kind of explanations children produced during their microgenetic sessions distinguished between improving and nonimproving children and were reliably related to improvement over time. In particular, children who gave more mentalistic explanations in response to general explanation requests were more likely to improve. In turn, children who showed more variety in their explanation responses were more likely to show patterns of FB explanation change associated with task improvement. Regression analyses further demonstrated that group differences in children's explanation patterns explained in part why children in the microgenetic group tended to improve but those in the comparison group uniformly did not: The microgenetic condition appeared to more readily elicit FB explanation patterns linked to task improvement.

Experimentally, there were two key differences between the microgenetic and comparison condition. First, children in the microgenetic condition generated explanations for every FB task they received, whereas children in the comparison condition did so for only half of their tasks. That microgenetic group children showed higher levels of improvement is consistent with the hypothesis that engaging in more frequent explanatory talk about FB events promotes theory of mind development. The greater amount and regularity of the explanatory talk these children engaged in may have been critical for generating mentalistic accounts of FB events and achieving success in predicting them.

Explanation variables did not fully mediate the relation between microgenetic condition and improvement, however, demonstrating that other factors contributed to group differences. The second major difference between the conditions lay in the spacing of children's microgenetic sessions. The microgenetic group's schedule of experiences was more spread out over time (i.e., fewer tasks per session over more sessions), but the comparison group's schedule was more condensed (i.e., more tasks per session over fewer sessions). Thus, another possible reason for group differences is that children learned about FB better under conditions of "distributed" rather than "massed" practice (Baddeley & Longman, 1978). That is, perhaps it was easier to process and consolidate new information about FB when it was presented in smaller chunks over a longer period of time, rather than in larger chunks over a shorter period of time. Our design did not allow us to disentangle the effects of having a more distributed schedule of FB experiences from the effects of having more opportunities to engage in explanation; we only know that when both of these features were present children showed higher levels of improvement. It seems likely that both explanation and experiences that provoke explanation-seeking, like the implicit feedback our tasks provided, contribute uniquely to developmental change. If opportunities for explanations were the only relevant factor, we might expect to see some level of improvement in the comparison group, who also had FB explanation experience, albeit to a lesser extent. In future work, we hope to tease apart the separate contributions of specific FB experiences and their organization over time.

Nonetheless, our results highlight the role of conceptually rich conversational experiences, including those involving explanation, in promoting theory of mind development. Why should these kinds of experiences be so crucial, over and above simple exposure and performance feedback? Consistent with a "theory–theory" view, we believe that theory of mind development involves a process of conceptual restructuring in which children rework their current understandings when challenged by new evidence. Children's engagement with rich conceptual information—in the form of adults' direct causal explanations, or in the form of joint explanation-seeking (which problematizes key issues in a way feedback alone cannot)—would seem to be central to this process. Such experiences support the emergence and consolidation of children's earliest concepts of belief, providing opportunities to learn and practice reasoning about these concepts, while also compellingly demonstrating their relevance and explanatory strength.

CONCLUSIONS

The understanding of theory of mind development afforded by previous cross-sectional and longitudinal research was one where we had snapshots of children's theories of mind before and after the transition to FB understanding. What microgenetic data, like those presented here, do, is flesh out our image of that transition itself, making possible a deeper understanding of the nature and path of change, as well as the mechanisms responsible for it. With regard to the path of change, our microgenetic group demonstrated improvements in FB understanding that occurred gradually over 9 to 10 weeks. From consistent failure on a variety of FB tasks at pretest, children progressively improved over 12 microgenetic sessions, culminating in the kind of consistent success on FB tasks typical of older preschool children. With regard to mechanisms, children's experiences producing explanations for FB events during the microgenetic sessions significantly accounted for their performance improvements, with more improvement occurring for children who showed increasing awareness of characters' thoughts and mistakes during the sessions. Further, this mentalistic orientation appeared to be more readily elicited in a microgenetic condition where children had more temporally distributed (rather than massed) FB experiences and more regular encouragement to seek explanations for these events.

A better understanding of the path of change in theory of mind development is an important new contribution in its own right. Coupled with increased insight into the sources of this change, our microgenetic findings carry significant implications for theoretical accounts of theory of mind development—which alternately describe development as primarily driven by innate maturational processes (e.g., Baron-Cohen, 1995; Leslie, 1994), or as depending critically on everyday experience in the social world (e.g., Gopnik & Wellman, 1994; Peterson & Siegal, 1995). Our data suggest that specific aspects of children's experiences, such as the particular kinds of explanatory conversations children engage in, are indeed related to developmental change. They also reveal for the first time how development unfolds in response to such experiences, revealing it to be gradual and protracted rather than rapid and complete. Such evidence is consistent with accounts of theory of mind development that place a high priority on children's experiences in the social world, and children's active attempts to make explanatory sense of those experiences, as central mechanisms of development.

ACKNOWLEDGMENTS

Jennifer Amsterlaw is now at the Institute for Learning and Brain Sciences, University of Washington.

This research was supported by NICHD Grant HD–22149, an NICHD graduate student traineeship, funds from the McDonnell Foundation, and NSF Grant SBE–0354453.

Parts of these data were presented at the October 2001 meeting of the Cognitive Development Society in Virginia Beach, VA, and at the April 2003 meeting of the Society for Research in Child Development in Tampa, FL.

We are grateful to the University of Michigan's Children's Centers and the parents and children who participated in this research. We thank Crystal Espinoza, Chien Tai Shill, and Amanda LeRoy for assistance with data collection and coding. We also thank Maureen Callanan and two anonymous reviewers for their thoughtful comments on a previous draft of the article.

REFERENCES

Appleton, M., & Reddy, V. (1996). Teaching 3-year-olds to pass false belief tests: A conversational approach. *Social Development, 5,* 275–291.

Baddeley, A. D., & Longman, D. J. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics, 21,* 627–635.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

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

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.

Clements, W. A., Rustin, C., & McCallum, S. (2000). Promoting the transition from implicit to explicit understanding: A training study of false belief. *Developmental Science, 3,* 88–92.

Dunn, J., Brown, J. R., & Beardsall, L. (1991). Family talk about feeling states and children's later understanding of others' emotions. *Developmental Psychology, 27,* 448–455.

Dunn, J., Brown, J. R., 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., Green, F., & Flavell, E. (1986). Development of knowledge about the appearance–reality distinction. *Monographs of the society for research in child development, 51*(1, Serial No. 212).

Flavell, J. H., & Miller, P. H. (1998). Social cognition. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language development* (5th ed., pp. 851–898). New York: Wiley.

Flynn, E., O'Malley, C., & Wood, D. (2004). A longitudinal, microgenetic study of the emergence of false belief understanding and inhibition skills. *Developmental Science, 7,* 103–115.

Gopnik, A., & Astington, J. (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., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press.

Guajardo, N. R. (1999, March). *Does discourse cause theory of mind development?* Poster presented at biennial meeting of the Society for Research in Child Development, Albuquerque, NM.

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.

Knoll, M., & Charman, T. (2000). Teaching false belief and visual perspective taking skills in young children: Can a theory of mind be trained? *Child Study Journal, 30,* 273–304.

Kuhn, D. (1995). Microgenetic study of change: What has it told us? *Psychological Science, 6,* 133–139.

Leslie, A. M. (1994). Pretending and believing: Issues in the theory of ToMM. *Cognition, 50,* 211–238.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development, 74,* 1130–1144.

Miller, P. H., & Coyle, T. R. (1999). Developmental change: Lessons from microgenesis. In E. K. Scholnick (Ed.), *Conceptual development: Piaget's legacy* (pp. 209–239). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

O'Neill, D., & Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology, 27,* 390–397.

Perner, J., Leekam, S., & Wimmer, H. (1987). Three-year-olds' difficulty understanding false belief: Representational limitation, lack of knowledge or pragmatic misunderstanding. *British Journal of Developmental Psychology, 5,* 125–137.

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

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

Sabbagh, M. A., & Callanan, M. A. (1998). Metarepresentation in action: 3-, 4-, and 5-year-olds' developing theories of mind in parent–child conversations. *Developmental Psychology, 34,* 491–502.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481–520.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology, 28,* 225–273.

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist, 46,* 606–620.

Siegler, R. S., & Svetina, M. (2002). A microgenetic/cross-sectional study of matrix completion: Comparing short-term and long-term change. *Child Development, 73,* 793–809.

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.

Taylor, M. (1996). A theory of mind perspective on social cognitive development. In R. Gelman & T. Au (Eds.), *Perceptual and cognitive development* (pp. 283–329). San Diego, CA: Academic.

Wahl, S. (2001, August). Three-year-olds' development in theory of mind: A microgenetic study. Paper presented at the 10th European Conference on Developmental Psychology, Uppsala, Sweden.

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

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72,* 655–584.

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.