Executive Function and Theory of Mind: Stability and Prediction From Ages 2 to 3.

Article *in* Developmental Psychology · November 2004

DOI: 10.1037/0012-1649.40.6.1105 · Source: PubMed

3 authors, including:

Dorothy J Mandell
University of Amsterdam

Jeffrey Lucas Williams
Portland VA Medical Center

Developmental Psychology
2004, Vol. 40, No. 6, 1105–1122

0012-1649/04/$12.00 DOI: 10.1037/0012-1649.40.6.1105

Executive Function and Theory of Mind: Stability and Prediction From Ages 2 to 3

Stephanie M. Carlson, Dorothy J. Mandell, and Luke Williams
University of Washington

Several studies have demonstrated a relation between executive functioning (EF) and theory of mind (ToM) in preschoolers, yet the developmental course of this relation remains unknown. Longitudinal stability and EF–ToM relations were examined in 81 children at 24 and 39 months. At Time 1, EF was unrelated to behavioral measures of ToM but was significantly related to parent report of children's internal-state language, independent of vocabulary size. At Time 2, behavioral batteries of EF and ToM were significantly related ($r = .50$, $p < .01$). Furthermore, EF (Time 1) significantly predicted ToM (Time 2), independent of several controls. A reciprocal relation (internal-state language → EF at Time 2) was nonsignificant with the controls included. Individual differences in EF were relatively stable.

One of the most striking features of early child development is the drive to decode one's social world. It is evident in children's attempts to make sense of others' behaviors by taking into account their mental and emotional states. This view of people as intentional agents is broadly construed as a theory of mind (ToM; for a review, see Flavell & Miller, 1998). A representational ToM assessed with measures like false belief tasks is not in place until age 5, although it is widely acknowledged that there are progenitors to a full-fledged understanding of mind (Perner, 1991; Wellman, 1990; Wellman, Cross, & Watson, 2001). Its development most likely draws on several interconnected skills and experiences. Correlates of ToM include language skills (e.g., de Villiers & de Villiers, 2000; Lohmann & Tomasello, 2003), mental-state talk in the home environment and between friends (e.g., J. Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991; Hughes & Dunn, 1998; Ruffman, Slade, & Crowe, 2002), sibling apprenticeship (e.g., Lewis, Freeman, Kiriakidou, Miridaki-Kassotaki, & Berridge, 1996; Ruffman, Perner, Naito, Parkin, & Clements, 1998), maternal "mind-mindedness" (Meins et al., 2002), an authoritative parental discipline style (e.g., Pears & Moses, 2003; Ruffman, Perner, & Parkin, 1999), and pretend play (Astington & Jenkins, 1995; Taylor & Carlson, 1997).

Among the cognitive developmental correlates of ToM, one in particular has been the subject of much theoretical and empirical investigation: executive function (EF). Despite growing evidence that individual differences in ToM are robustly correlated with EF skills in the preschool period (e.g., Carlson & Moses, 2001; Hughes, 1998a; Perner & Lang, 2000; Russell, Jarrold, & Potel, 1994), the developmental course of this association is not well understood. We report on a longitudinal study designed to address the nature of relations between young children's emerging understanding of the mind and executive functioning from 2 to 3 years of age. In what follows, we provide an overview of links between EF and ToM and then describe the aims and challenges of the present study.

Executive Function and Theory of Mind

Executive functioning refers to higher-order self-regulatory cognitive processes, including the control of attention and motor responses, resistance to interference, and delay of gratification, that are typically associated with operations of prefrontal cortex (Duncan, 1986; for a review, see Zelazo, Carter, Reznick, & Frye, 1997). EF is often assessed using neurocognitive tasks such as Luria's Hand Game (Luria, Pribram, & Homskaya, 1964). In this game, children are given a set of imitation trials in which they are to mimic the hand position of the experimenter (closed fist or pointed finger) as well as a set of anti-imitation trials in which they are required to make the opposite gesture. These anti-imitation trials are considerably more difficult for young children and for adults with prefrontal brain lesions, and reaction time is slowed (see Meltzoff & Decety, 2002, for a review of the neuroscience basis of imitation). Performance on EF tasks may be moderated by working memory capacity (Engle, 2002; Roberts & Pennington, 1996). For example, in the Hand Game, the ability to hold the rules in mind is critical to task success.

There are several reasons to suspect that EF and ToM are fundamentally linked in development (see Carlson & Moses, 2001; Moses & Carlson, in press). First, at the same time when pre-

Stephanie M. Carlson, Dorothy J. Mandell, and Luke Williams, Department of Psychology, University of Washington.

This research was generously supported through a gift to the University of Washington Center for Mind, Brain, and Learning from the Talaris Research Institute and the Apex Foundation, the family foundation of Bruce and Jolene McCaw. Additional support was provided by the Royalty Research Fund of the University of Washington.

We are grateful to the families who participated. We also thank Andrew Meltzoff for many helpful suggestions, Liliana Lengua for statistical consultation, and Craig Harris, who provided training on the Discrepant Desires task for preschoolers. Several undergraduates assisted with this research: Benjamin Berres, Kima Brown, Tamara Canero, Addy Clark, Caron Cosser, Briana Coyle, Heather Fuqua, Jamie Gum, Jenee Hoard, David Kurihara, Meg Lemke, Andrea Molloy, Tiffany Olsen, Suzanna Ramirez, Elissa Ross, Tiffany Wang, Valerie Weiss, Emily Whitish, Eylul Wintermeyer, and Valerie Wood.

Correspondence concerning this article should be addressed to Stephanie M. Carlson, Department of Psychology, University of Washington, Seattle, WA 98195–1525. E-mail: carlsons@u.washington.edu

schoolers have great difficulty taking someone else's perspective, as in false belief tasks, they also have obvious shortcomings in self-control. They have a very hard time resisting temptation, controlling attention, and keeping their actions and utterances in check. The developmental timetable of major growth in self-regulatory skills parallels that of ToM (e.g., Kochanska, Coy, & Murray, 2001). Second, common neural underpinnings have been posited in both EF and ToM, although different areas of the frontal lobes tend to be implicated on different measures (e.g., Gallagher & Frith, 2003; Siegal & Varley, 2002). Third, impairments in certain aspects of both EF and ToM have been implicated in the developmental disorder of autism (Hughes & Russell, 1993; Ozonoff, Pennington, & Rogers, 1991), although the specificity of executive dysfunction in autism has been questioned (e.g., Griffith, Pennington, Wehner, & Rogers, 1999). Finally, a process analysis of what is required for successful reasoning about mental states would suggest that *some* level of executive skill is needed. For example, in the standard, unexpected-location false belief task (Wimmer & Perner, 1983), young children who have relatively poor executive control may find it extremely difficult to resist pointing to the current, highly salient location of the object in question. Moses (2001) has referred to this as an expression account of the EF–ToM relation: Executive deficits limit children's ability to express their ToM competence. Indeed, when the inhibitory demands of a deceptive pointing task were reduced by having children use a symbolic indicator (picture or arrow) rather than finger pointing, even 3-year-olds were able to deceive above chance levels (Carlson, Moses, & Hix, 1998; Hala & Russell, 2001).

Beyond the executive demands of standard ToM tasks, however, some investigators have speculated on a more fundamental connection between EF and ToM; that is, a certain level of self-regulation is critical for the very *emergence* of mental-state concepts (Moses, 2001). In this view, children must be able to disengage from salient but misleading stimuli in the environment and suppress their own potent representations of events before accurate reflections on the mental states of self and other can be achieved (Russell, 1996). Perner (1998; Perner & Lang, 2000) has proposed an alternative conceptualization that ToM must be acquired before children can reflect on and subsequently monitor and control their own behavior (see also Carruthers, 1996). Frye (2000) and Frye, Zelazo, and Burack (1998) maintained a third position on the relation between EF and ToM. They proposed that a cognitive milestone in the preschool period, the domain-general ability to reason about complex problems involving selective attention, gives rise to improved performance on both executive control and ToM tasks.

There is evidence from correlation studies that is consistent with each of these accounts. Carlson and Moses (2001) found that ToM in 3- and 4-year-olds was significantly correlated with EF measures requiring children to inhibit a dominant response and initiate a subdominant response (conflict tasks) and measures requiring them to wait before making a response (delay tasks). Other studies have reported similar findings (e.g., Frye, Zelazo, & Palfai, 1995; Hughes, 1998a, 1998b; Hughes, Dunn, & White, 1998; Perner & Lang, 2000). It is interesting that Carlson and Moses (2001) found that correlations were consistently higher for conflict tasks than delay tasks. They attributed this result to the combination of inhibitory and working memory requirements of conflict tasks. In contrast, delay tasks have high inhibition demands but relatively low working memory demands. Subsequent work confirmed that inhibition and working memory together appear to play a critical role in ToM performance (Carlson, Moses, & Breton, 2002; Hala, Hug, & Henderson, 2003).

Supporting emergence accounts of the EF–ToM relation, recent studies have shown that it is more than a simple matter of inhibition *in* the ToM tasks. Perner, Lang, and Kloo (2002) reported that EF was related to a false belief explanation task in which the inhibitory demands were reduced compared with a standard prediction task. Similarly, Moses and Carlson (in press) reported a strong relation between individual differences in EF and *nonstandard* ToM tasks in which there was no clear prepotent response bias (e.g., a sources-of-knowledge task). They concluded that an underdeveloped executive/regulatory system could constrain the emergence and expression of concepts about the mind.

Our understanding of EF–ToM relations is far from complete, however. For example, the developmental ordering driving the relation is debated (Perner & Lang, 2000; Russell, 1996). In an important study of these issues, Hughes (1998b) tested 50 children on measures of EF and ToM at a mean age of 3 years 11 months and again at 5 years 0 months. Performance on a conflict EF task significantly predicted ToM 1 year later, independent of age, verbal ability, and earlier ToM scores. There was no evidence of a reciprocal relation (ToM predicting EF). Hughes concluded that even normal variation in EF might have consequences for young children's developing "mentalizing" ability. More recently, Kloo and Perner (2003) conducted a training study of false belief and a common measure of EF, the Dimensional Change Card Sort (DCCS; Frye et al., 1995), with 44 three- and four-year-olds. Explicit training on the DCCS improved false belief performance and vice versa. Kloo and Perner concluded that although their results provided further support for a functional dependency between EF and ToM, they did not shed light on the processes underlying this interdependence and its developmental course. Further longitudinal work is needed to gain a more complete understanding of the nature and origins of EF–ToM relations.

Goals of the Present Study

The purpose of the present study was to examine the developmental associations between EF and ToM by beginning earlier in development than previous studies have. Specifically, we decided to focus on the 3rd year of life, a period when EF and potential "precursors" to ToM are rapidly developing and when individual differences in these skills emerge. Starting early in development may help us learn when an EF–ToM relation first appears and trace its stability over time. In addition, the design of this study makes it possible to examine the coherence of EF and ToM constructs in early childhood. This research represents a first step in establishing predictors of children's executive control and mental-state understanding from 24 months of age.

There are challenges to beginning early in development, however. One is that standard ToM measures like the false belief task are generally too difficult for children younger than 3 years. Several investigators have proposed that ToM develops gradually and encompasses conceptual changes that precede an understanding of belief. For example, Gopnik, Slaughter, and Meltzoff (1994) suggested that toddlers' understanding of differing visual perspectives could promote a more advanced awareness of differing "points of view." Related arguments have been made for toddlers' ability to read into the mental states of intention (Meltzoff, 1995) and desire (Repacholi & Gopnik, 1997; for a review of "theory theory," see Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994). Similarly, Perner (1991) proposed that the conceptual shift occurring in children's representation of mental states goes hand in hand with more general advances in their representational understanding. Therefore, to address the obstacle of measuring ToM in the young participants in our study, we assessed children's broader understanding of mental representation, including seeing, wanting, pretending, believing, and distinguishing representations from reality with objects (e.g., appearance–reality tasks).

A second challenge was to obtain performance variability on both EF and ToM at each age. Given a dearth of established measures of these constructs that can be administered in a parallel fashion at ages 2 and 3 (with substantial variability at each of those ages), we were unable to carry out a repeated measures design. To address this difficulty, we selected measures that were age-appropriate and therefore unlikely to produce floor and ceiling effects. Our aim was to include a set of EF and ToM tasks at each age that would be coherent and draw out individual differences. Children's relative standing on performance in each domain was then analyzed both concurrently and over the 15-month span.

The study began with an assessment of EF and ToM in a sample of 81 typically developing toddlers (24 months). Their parents completed questionnaires regarding child temperament, mental-state word usage, and vocabulary level. Approximately 1 year later, the families were recontacted and children (at 39 months) were assessed on age-appropriate EF and ToM task batteries as well as a measure of verbal ability. Parents again completed questionnaires about their child's current functioning in these areas. We then analyzed (a) the coherence and stability of each construct separately and (b) the concurrent and longitudinal relations between EF and ToM. Given prior research on effortful control in early childhood (Kochanska, Murray, & Harlan, 2000), we hypothesized that individual differences in EF would be stable over time and, at least at the age 3 assessment, would be positively related to ToM, in keeping with the results of numerous other studies of preschoolers. However, we made no specific prediction about the stability of individual differences in ToM because to date no other study has examined ToM behavioral tasks in toddlers followed longitudinally. Similarly, we set out to test both major theories about the functional link—that ToM is dependent on EF (e.g., Moses & Carlson, in press; Russell, 1996) and that EF is dependent on ToM (e.g., Perner & Lang, 1999)—as well as the possibility of a bidirectional relation.

Method at Time 1 (24 Months)

Participants

The sample consisted of 81 children from the metropolitan Seattle area who participated in a longitudinal study at 24 and 39 months of age. From the original group of 118 enrolled families, 94 (80%) returned for the assessment at Time 2. Five families moved out of the area, 5 were not interested in participating, and 14 families could not be contacted. Of the 94 children who participated at Time 2, 6 children did not return for the second session and 7 additional children refused a large number of tasks. Thus, the final sample included 81 children (40 boys and 41 girls). At Time 1, the mean age was 24 months (SD = 1.41).[1] We recruited families by telephone from a database containing vital statistics records. The sample was predominantly White. Most families included two parents. The median age of mothers was 37 years (fathers = 38 years). The modal education level for both parents was a college degree. Parents were paid $10. Children received small snacks and prizes at each session.

Procedure

Children were tested in a laboratory playroom for two 45-min sessions spaced 1 week apart (mode = 7 days, M = 7.02 days, SD = 1.62). The same female experimenter and one of four assistants conducted the sessions. The accompanying parent was in the testing room, seated behind the child in a separate chair, and was given instructions to remain neutral. All sessions were videotaped.

We administered batteries of tasks designed to assess early EF and ToM abilities. Because our major interest in this study was individual differences, all tasks were presented in a fixed order, with the EF and ToM tasks interspersed throughout the two sessions. The tasks are described briefly in the following section.[2] We double-coded 31% of the sessions (n = 25). Scoring was reliable (Cohen's κs $\geq$ .80; agreement on categorical variables $\geq$ 90%; all reaction times within 1 s). Discrepancies were resolved by a third coder who was unaware of the hypotheses of the study.

Measures

The measures included batteries of EF and early ToM tasks as well as a set of parent-report measures.

Executive Function

The EF battery consisted of five tasks designed to measure the ability to control prepotent responses.[3]

Reverse Categorization. Children were required to sort big and little blocks according to their size into big and little buckets and then to reverse this categorization scheme so that the big blocks would go into the little bucket and vice versa. The experimenter demonstrated six preswitch trials (three of each size), stating the rule on each trial (e.g., "Here's a *big* block, it goes in the *big* bucket"), and then asked children to sort the remaining six blocks. The experimenter then emptied the buckets and introduced the test trials, slowly saying, "Okay, now let's play a silly game. Let's put all the *big* blocks in the *little* bucket and put all the *little* blocks in the *big* bucket." The experimenter restated the rule and identified the size of the block before each trial. There was no feedback on the 12 test trials. Scores were the proportion of correctly sorted blocks (postswitch). This task is similar to Perner and Lang's (2002) Reversal Shift task for older children, which uses cards with pictures rather than three-dimensional objects (see also Brooks, Hanauer, Padowska, & Rosman, 2003). These single-dimension switch tasks are easier (and therefore more age-appropriate for our study) than analogous sorting tasks that use two integrated dimensions (Brooks et al., 2003).

Multilocation Search. This task is a modified A-not-B task (see Diamond, 1988) developed by Zelazo, Reznick, and Spinazolla (1998) for use

[1] We confirmed that the sample of children who returned for the Time 2 assessment was not significantly different from those who did not return. Analyses comparing these groups (t tests) on Time 1 age, sex, verbal ability, and the EF and ToM measures showed no differences (ps $\geq$ .54). These results suggested that attrition from Time 1 to Time 2 was random.

[2] Complete descriptions of the tasks are available from Stephanie M. Carlson.

[3] One additional EF task (Distracting Box) was piloted but dropped because of ceiling effects.

with toddlers. It requires children not only to track changes in the location of a hidden reward but also to associate a symbol with the correct location. Procedures followed specifications by Zelazo et al. The experimenter introduced a wooden apparatus having five drawers aligned in a row. In the warm-up, she attached a symbol (a green diamond shape) by a string to the center drawer, placed a snack treat (e.g., a goldfish cracker) in the drawer (from the back, accessible only to the experimenter), and demonstrated pulling on the symbol on the child's side to open the drawer and reveal the treat. Children practiced retrieving the snack in this manner. Next, the experimenter attached different symbols to three drawers, placed a snack in the middle drawer, announced its location, placed a felt cover over the apparatus, and finally invited children to find the snack. She repeated this procedure until children successfully retrieved the snack three times in a row. For the postswitch trials, the process was repeated except the snack was placed—in full view of the child—in a *different* drawer and the experimenter imposed a 10-s delay before inviting children to get the snack. Postswitch trials continued until children were able to retrieve the snack without perseverating (i.e., first going to the previously rewarded symbol). We recorded the number of postswitch perseverative errors (reverse scored in analyses).

Shape Stroop. Similar to a Stroop task, this task assesses children's ability to focus on a subdominant perceptual feature. Following Kochanska et al. (2000), we showed children individual colored line drawings of big and little apples, oranges, and bananas and asked them to point to each as the experimenter named them. The experimenter then replaced these cards with three cards depicting a little fruit nested in a different large fruit (e.g., a little apple in a big banana). The experimenter asked children to point to each of the little fruits (e.g., "Show me the *little* apple"). Responses were coded as follows and averaged across trials: 0 = points to the large fruit; 1 = points to the large fruit but self-corrects; and 2 = points to the little fruit.

Snack Delay. The following two tasks were adapted from Kochanska et al. (2000). The experimenter placed a snack treat (e.g., a goldfish cracker) under an inverted transparent cup and, after a brief warm-up, instructed children to wait until she rang a bell before getting the snack (four trials, delays of 5, 10, 20, and 15 s, respectively). The experimenter restated the rule before each trial. We recorded the time children waited before retrieving the snack and averaged it across trials (maximum possible score = 12.50 s).

Gift Delay (bow). The experimenter presented a large shopping bag containing a shiny wrapped gift box and told children she had a present for them. The experimenter looked in the bag and said, "Uh oh! I forgot the bow! Let me go get it. But let's play a game again. Sit here and don't touch the present until I come back with the bow, okay?" The experimenter and the assistant left the room for 180 s or until the child opened the gift. Following Kochanska et al. (2000), we gave children a touch score ranging from 1 to 5, based on the degree of restraint they exhibited (5 = *no touch to bag or gift*).

Theory of Mind

The ToM battery consisted of four tasks designed to measure early understanding of mental states.

Intentions. In this task, adapted from Meltzoff (1995), children were presented with three different scenarios in which the experimenter's intended actions on objects remained unfulfilled. We selected props with actions that children this age did not perform spontaneously in pilot testing. The actions were pulling apart a dumbbell-shaped toy, driving a toy truck into the center of a toy horseshoe, and placing a plastic stick through an opening at the top of a toy music ball. The experimenter first presented the toy sets individually. Children were allowed to play with each set for 1 min. She then demonstrated each of the scenarios in turn while maintaining a neutral facial expression. The experimenter repeated the unsuccessful action three times (e.g., driving the truck to the left of the opening in the horseshoe, then to the right, then repeating twice). No verbal instructions were given except "watch this" to orient children to the experimenter's actions during demonstrations. Then the experimenter gave children the stimuli again, one at a time, with no instruction. Children received credit if they pulled the dumbbell toy apart, drove the truck into the center of the horseshoe, or placed the stick in the opening on top of the music ball within 20 s. If children had performed the target action during the initial exploratory phase (which occurred infrequently), their score on the corresponding test trial was disqualified. Thus, proportions, calculated as the number of trials in which children performed the adult's intended act versus the actual act or another act, divided by the number of qualifying trials, were used in analyses.

Discrepant Desires. Following Repacholi and Gopnik (1997), in this task we presented children with two bowls on a tray. One contained a goldfish cracker or pretzel, and the other, a small floret of broccoli. The experimenter invited children to eat whichever snack they wanted for 45 s. This snack was judged to be the preferred choice of the child. The experimenter pulled the bowls toward her and out of reach of the child, emptied them, and replaced both snacks. Then she picked up each bowl in turn, looked at the food, and made a facial expression of disgust while saying a long "Eeew" in reference to the preferred snack of the child. Next she made a happy face and said a long "Mmm" in reference to the child's nonpreferred snack (usually broccoli). Each expression lasted 5 s. The experimenter repeated this sequence once and then held out her hand midway between the bowls and requested some food ("Can you give me some?") while moving the tray toward the child. The experimenter made up to four requests and moved on if she was not offered anything. This process was repeated for a second trial with cold black coffee and a juice box. Unlike the original task developed for 18-month-olds (Repacholi & Gopnik, 1997), the experimenter did not taste the food choices. Children received proportion scores for correct offers (corresponding to the experimenter's preference).

Visual Perspective. We developed this task to assess children's early understanding of visual perspective, that is, that one must have visual access to objects to see them and share the experience of them. The experimenter brought the child to one corner of the room, and the parent was directed to sit on the floor in the opposite corner. After a brief warm-up, children were instructed to show their parent a variety of toys across six test trials. The experimenter positioned children so that their backs were to their parent, handed them a toy, and said, "See this? Show it to mommy[daddy]." Children then turned around and walked to the parent to show her or him the toy and then were called back to the experimenter's corner of the room. A research assistant seated near the parent whispered directions between each trial. Parents were instructed to occlude their vision by: (1) closing their eyes; (2) placing their hands over their eyes; (3) pulling a hat down over their eyes; (4) orienting themselves so that a screen was directly in front of them; and (5) turning their back to the child. The sixth and final trial involved a snow globe containing a scene that could only be seen when viewed from one side and so would need to be turned around to face the parent. We coded whether children attempted a correction, that is, acknowledged that the parent could not see the object and adjusted themselves, the parent, or the toy so that the parent had clear sight of the object (e.g., removed the hat and held up the object in front of the parent's eyes). Children received a score from 1 to 5 on each trial: 1 = no show or dropped toy in parent's lap and walked away; 2 = held toy near parent but no attempt at correction; 3 = partial correction; 4 = full correction but did not show the toy; 5 = full correction and showed the toy. We used average scores across trials in data analysis.

Comprehension of Pretense. This task, adapted from Harris and Kavanaugh (1993), assesses toddlers' ability to follow along with and reenact a pretense representation. The experimenter placed four blocks on the table, including two yellow semicircle blocks to the right of the child and two brown square blocks to the left. The experimenter introduced a monkey puppet and announced, "Monkey is hungry. I'm going to feed monkey some banana." Then she pretended to feed a yellow block to the monkey (holding it to its mouth and making "yum yum" sounds). Next the experimenter said, "Monkey wants some more. Feed monkey some banana," and handed the child a yellow block. After children fed the monkey, the experimenter introduced a toy horse. She announced that the horse was also hungry but wanted some cake. The experimenter used the brown block to feed the horse and encouraged the child to also feed the horse some cake. The animals were removed and the blocks were reset on the table. For the test trials, the experimenter introduced a toy cow, pig, dog, and cat in turn and stated that the animal wanted banana or cake (two of each). We recorded the number of correct offers out of four. No feedback was given on test trials.

Parent Questionnaires

At Session 1, we gave parents three questionnaires assessing their child's vocabulary, use of internal-state words, and overall temperament. Three families did not return any of the questionnaires.

MacArthur Communicative Development Inventory (CDI). Parents completed the MacArthur CDI Toddler Form (Fenson et al., 1993). They marked which words their child was currently saying from a set of over 600. Total scores were used in analyses.

Internal States Language Questionnaire (ISLQ). Parents were given the ISLQ, a questionnaire designed to measure children's production of language conveying mental, emotional, motivational, and physiological states (Bretherton & Beeghly, 1982). Parents indicated which words their child was currently saying from a list of 77 items (e.g., "see"). For all words produced, parents were asked to note the context—whether referring to self, others, or a picture or representation (e.g., a character in a movie or book)—and to provide a specific example for each particular usage. Total scores were used in analyses.

Toddler Behavior Assessment Questionnaire—Revised (TBAQ–R). This temperament questionnaire was developed by Goldsmith (1996) and revised by Rothbart (personal communication, March 2000). It includes 100 Likert-scale items comprising 13 subscales. Parents rated each item on how well the description fit their child, on a scale from 1 (*extremely untrue*) to 7 (*extremely true*). We were most interested in the subscales of Inhibitory Control, Attentional Focus, Positive Anticipation, and Perceptual Sensitivity for the present study.

Method at Time 2 (39 Months)

Participants

The same 81 families participated when the children were 39 months of age (M = 39 months, SD = 0.91; 40 boys and 41 girls). Parents received $14, and children were given snacks and small prizes.

Procedure

We tested children individually in two videotaped 45-min sessions approximately 1 week apart (mode = 7 days, M = 9.73 days, SD = 5.78). The EF and ToM tasks were presented in a fixed, interspersed order. The same male experimenter tested all children. Parents observed the sessions from an adjacent room. We double-coded 33% of the sessions (n = 27). Scoring was reliable (Cohen's κs $\geq$ .80; agreement on categorical variables $\geq$ 90%; all reaction times within 1 s). Discrepancies were resolved by a third coder.

Measures

Measures included a verbal ability test, an EF battery, a ToM battery, a mental state control task, and parent-report measures.

Verbal Ability

The experimenter administered the Peabody Picture Vocabulary Test—Third Edition (PPVT-3; L. M. Dunn & Dunn, 1997) following the standard protocol. Raw scores (ceiling item − number of errors) were used in data analyses.

Executive Function

Reverse Categorization. The experimenter introduced children to six pairs of toy zoo and farm animals. Each pair consisted of a large ("mommy") and small ("baby") model of the same animal (polar bears, tigers, cows, pigs, horses, and elephants). Then children were shown two differently colored buckets, one with a picture of a woman ("mommy") attached to it, the other with a picture of a baby. The rest of the procedure was exactly parallel to that at Time 1 except that the experimenter reminded children of the rules before the first and seventh trials, rather than on every trial. Scores were the proportions of correctly sorted animals.

Grass/Snow. Following Carlson and Moses (2001), we had children place their hands on top of felt hand cut-outs on a table and asked them to point to a white card when the experimenter said "grass" and to a green card when he said "snow." After a brief warm-up, the experimenter administered 16 test trials (8 of each verbal cue) in a fixed random order. Proportion scores were used in data analyses.

Snack Delay (Choice). This task was adapted from Mischel, Shoda, and Rodriguez (1989). Children selected a preferred snack (e.g., fruit-flavored sugar cereal), and the experimenter put two of the snacks in one bowl and 10 in another bowl. He asked children which pile they would like. (All children selected the larger pile.) The experimenter told the children he needed to leave the room to do some work and that if they did not leave their seats or eat the snack, they would be given the large pile (pointing to it) upon his return. However, if they did not wait, they were told they could only have the smaller pile (pointing to it). The experimenter confirmed children's understanding of the rules and then exited, waiting up to 60 s or until the child ate a snack. We scored difficulty of delay, including the length of time to the first touch of the snacks and total number of touches (reverse scored), as well as total waiting time. All three variables were significantly intercorrelated (rs $\geq$ .66, ps $<$ .001), and so were standardized and averaged.

Bear/Dragon. In this task, adapted from Reed, Pien, and Rothbart (1984), children had to alternate between performing and suppressing commanded actions. After a brief warm-up, the experimenter first introduced a "nice bear" puppet (using a soft voice) and said they should do what the bear told them to do, and then he introduced a "naughty dragon" puppet (using a gruff voice) and said they should *not* listen to it. Practice trials (up to six) and a verbal rule check followed (with corrective feedback). There were 10 test trials, alternating between the bear and dragon commands. The experimenter issued a rule reminder halfway through the 10 trials. Responses were scored from 0 to 3 for each trial. A full commanded movement was scored as 3 when given by Bear and as 0 when given by Dragon; a partial commanded movement was scored as 2 for Bear and as 1 for Dragon; a wrong/different movement was scored as 1 for Bear and as 2 for Dragon; and no movement was scored as 0 for Bear and as 3 for Dragon. Bear and Dragon scores were summed. Test trial scores were significantly related to the number of practice trials required for the Dragon (reverse scored), $r(69)$ = .75, p < .001, and so were aggregated to form composite Bear/Dragon scores, following Carlson and Moses (2001).

Hand Game. We followed Hughes's (1998b) procedure, originally adapted from Luria et al. (1964). In the warm-up, the experimenter verified that children could point a finger and make a fist as the experimenter modeled these hand movements. Six imitation trials followed in which children were instructed to make the same shape with their hand as the experimenter (fist or pointing finger). Next, the experimenter introduced the anti-imitation trials. He explained that the children were now to make the shape opposite to the one he made (e.g.,"When I point my finger, you make a fist"). The experimenter administered four practice trials with alternating gestures. Fifteen test trials (8 fist and 7 pointing finger) followed in a fixed random order. Scores were the proportion of trials correct.

Tower Building. The following two tasks were adapted from Kochanska, Murray, Jacques, Koenig, and Vandegeest (1996). The experimenter presented children with 20 blocks and told them they should take turns with him in building a tall tower. After a brief demonstration of turn-taking, the experimenter let children begin building the tower. He did not take a turn unless children explicitly communicated that they were giving him a turn (e.g., verbally or by handing him a block). Children received a point for each block placed by the experimenter. Following Kochanska et al. (1996), we subtracted points if children knocked down the tower prematurely (−5) and added points if they carefully removed the blocks after finishing the tower (+5). We converted the point total to a proportion of all blocks used (ideally, .50). The mean of the two trials was used in analyses.

Whisper. This task called for tempering a verbal response. To warm up, the experimenter asked children to whisper their names. Then he explained they would play a game of whispering the names of cartoon characters presented on cards. Six of the 10 characters were easily recognized by most 3-year-olds (e.g., Winnie the Pooh), and four were relatively unknown (e.g., Petunia). The experimenter whispered throughout and gave a rule reminder halfway through. Following Carlson and Moses (2001), we gave a score of 0 for shouting, a score of 1 for a normal voice or mixed shouting and whispering, and a score of 2 for whispering. We excluded trials in which children gave no answer. Average scores were used in analyses.

Gift Delay. The experimenter announced he had a present. He asked children to sit in a chair facing away from him and not to look while he wrapped their gift so that it would be "a big surprise." The experimenter noisily wrapped a present in a standardized manner for 60 s. We recorded a peeking score (0 = turn around to peek, 1 = peek over the shoulder, 2 = no attempt to peek) as well as the latency to peek over the shoulder and the total number of peeks (reverse scored). These scores were highly correlated (rs $\geq$ .78) and thus were standardized and averaged.

Theory of Mind

We administered six tasks designed to assess individual differences in young 3-year-olds' ability to understand mental states and discrepancies between representations and reality.[4]

Pretend–Reality. Following Frye et al. (1995), the experimenter stipulated a pretend identity of four objects and asked children to differentiate between the pretend and real identities. The four test items (and pretend identities) were string (snake), block (car), pencil (hammer), and sunglasses (telephone). While pretending (e.g., wiggling the string along the table), the experimenter asked children what they were *pretending* the object was (snake), and what it *really* was (string). The order of pretend and reality questions alternated across items. Scores were the proportion correct on the eight test questions.

Discrepant Desires (preschool). We used a variant of the toddler procedure used by Repacholi and Gopnik (1997) that has been tested successfully with preschoolers (A. N. Meltzoff, personal communication, August 2001). The experimenter presented a pair of snacks and told children they could try them. Each pair comprised an unusual choice (e.g., a vegetable) and a typical choice (e.g., a cookie). There were four pairs: celery and animal crackers, cauliflower and raisins, pickles and fruit cereal, and lettuce and teddy bear cookies. The experimenter observed children tasting the foods to distinguish their preferences. (When there was no clear preference, he assumed the child preferred the typical choice.) Next, the experimenter tasted both foods, displaying negative affect (disgust, grimace, "Eeew! I tasted [food label]") for the child's preference and positive affect (smile, raised brow, "Mmm! I tasted [food label]") for the child's nonpreferred snack. He then extended a hand between the snack plates and asked, "Can you give me some?" Scores reflected the proportion of correct offers to the experimenter.

Visual Perspective Taking—Level 1. Following Flavell, Everett, Croft, and Flavell (1981), we assessed children's knowledge about visual perception. Level 1 requires an understanding that someone else can be viewing an object different from the one viewed by oneself, whereas Level 2 requires the additional insight that the same object can be viewed differently by self and other. The experimenter showed children a card with a drawing of a cat on one side and a dog on the reverse side and asked them to identify the animals. He then held up the card so that the cat was facing the child and the dog was facing him, and asked children to identify which animal they saw and which animal the experimenter saw. He then flipped the card and repeated the procedure for a total of four questions on the dog/cat picture. Next, the experimenter placed an 8½ × 11 in. drawing of a turtle flat on the table, oriented to the children's perspective. The experimenter pointed to and labeled the turtle's feet and shell and then asked children to label these parts. Then he placed a plain white card on top of the turtle drawing and perpendicular to the table in order to separate the top and bottom portions of the drawing. The experimenter asked the children to identify which part of the turtle *he* could see. The experimenter then rotated the drawing 180°and repeated the question. Scores for the six questions were averaged.

Visual Perspective Taking—Level 2. In this task from Flavell et al. (1981), children were shown drawings and asked to determine how the drawings looked from another's perspective, the "right way" or the "wrong way." To warm up, the experimenter asked children whether a cup (held upright, then inverted) was held the right way or the wrong way, to ensure understanding of these terms. Then the experimenter presented three animal pictures (turtle, bird, and pig, in turn) and asked children how the animal looked *to the experimenter*. Each card was presented twice—right side up and upside down from the child's perspective—in a fixed alternating order. Next, the experimenter showed children a picture book right side up and then upside down and asked them to describe how it looked from their perspective. Then the experimenter stated that for the rest of the game, they should describe how the book looked from *his* perspective. Three test trials followed—twice the wrong way and once the right way from the experimenter's point of view. Scores reflected the proportion of correct responses for the animal and book trials (out of nine).

Appearance–Reality. Following Flavell, Green, and Flavell (1986), we showed children two objects that had misleading appearances. One object involved a discrepancy between its actual and apparent identity (a sponge shaped and painted to look like a rock). The second object involved a discrepancy between its actual and apparent color (a picture of a red castle that appeared black when behind a green filter). The experimenter first showed children both aspects of the sponge/rock and then asked what the object looked like and what it really was. Then he repeated this procedure for the red/black castle. Children received a score of 1 for each task by correctly answering both the appearance and reality questions. Scores on the two trials were averaged.

Contents False Belief. This task was adapted from Perner, Leekam, and Wimmer (1987) and Gopnik and Astington (1988). Children were shown a Band-Aid box and asked to tell what they thought was inside. (All said "Band-Aids.") Then the experimenter opened the box to reveal that it actually contained crayons. The experimenter closed the box and asked, "When you first saw this box, before we opened it, what did you think was inside, Band-Aids or crayons?" Children were then shown a doll, Bert, and told he had never looked inside the box. The experimenter asked the children what Bert thought was inside the box. Finally, children were asked to identify the true contents of the box. A correct answer on the control

[4] One additional ToM task (Discrepant Beliefs; Wellman & Bartsch, 1988) was included but later dropped because children were confused, as evidenced by inconsistent performance across trials and on the control questions.

question was necessary for credit on the "self" and "other" test questions. Proportion scores were recorded.

Non-Mental-State Change. This task, used by Gopnik and Astington (1988), mirrors the format of a contents false belief task but avoids language referring to mental states. We included it to control for simple inhibition and memory problems in examining the relation between EF and ToM. Children were shown an unmarked box containing a toy cat. Then the experimenter removed the cat and placed a toy rabbit inside. He closed the box and asked children what was currently inside, and then asked what was inside it before they opened it. Children were given a score of 1 if they answered both questions correctly (and a score of 0 otherwise).

Parent Questionnaires

Parent questionnaires were distributed at Session 1 and collected at Session 2. Six families did not return any questionnaires.

Children's Behavior Questionnaire—Short Form (CBQ). The CBQ was included to obtain a global assessment of children's temperament, measured through 94 Likert-scale questions covering 16 subscales (Putnam & Rothbart, 2001). In this investigation we were most interested in the subscales of Inhibitory Control, Attentional Focus, Impulsivity, and Perceptual Sensitivity.

Rules Questionnaire. Parents were given the Rules Questionnaire adapted from Smetana, Kochanska, and Chuang (2000). It measures rules and expectations parents have for their 3-year-old child. Twenty-six possible rules or expectations were listed (e.g., "putting toys away"). We asked parents to rate on a 5-point scale how important each rule was to them. Total scores were calculated.

Results

In the following sections, we describe results on the coherence and stability within the EF and ToM constructs and then the nature of the relations between them at 24 months, at 39 months, and across this age span.

Executive Function

Descriptive statistics of children's performance on the EF tasks at 24 and 39 months are shown in Table 1.[5]

Data Reduction

We first conducted principal-components analyses to inform the creation of battery scores on EF at each time point. The first unrotated principal component (FUPC) accounts for the maximum amount of variance of the measured variables. Item loadings reflect the correlation between a particular measure and the overall component. Loadings of .30 or above are typically considered acceptable (Tabachnick & Fidell, 1983). At 24 months, all five EF tasks had loadings above this threshold on the FUPC, as shown in Table 2. On the basis of these results, we averaged the standardized (z) scores on the five tasks to achieve equal weighting and computed an EF Composite (Time 1) score for each participant. However, because not all EF tasks were significantly intercorrelated (see Table 3) and because Cronbach's alpha was modest (.55), we also report on the individual tasks in the major analyses. (These and all subsequent tests were two-tailed.)

At 39 months, all eight tasks had item loadings $\geq$ .50 (see Table 2) and high internal consistency (Cronbach's α = .77). Therefore, we standardized and averaged scores on these eight items to create EF Composite (Time 2) scores. The intercorrelations among these tasks are provided in Table 4. Our method of data reduction is consistent with that of Kochanska et al. (2000), who reported lower coherence of EF measures at 22 months than at 33 months (αs = .42 and .77, respectively). Rushton, Brainerd, and Pressley (1983) advocated aggregation across measures for greater psychometric precision. When multiple indicators of a construct are included, the variance idiosyncratic to any particular task is cancelled out across measures. Controlling for age and intelligence in subsequent analyses also helps to lessen concerns that task intercorrelations are attributable to shared method variance or difficulty levels (e.g., Carlson & Moses, 2001; Hughes, 1998a).

Developmental Change

Reverse Categorization is the only EF task that was directly comparable across the two time points. As can be gleaned from Table 1, children showed significant improvement on this task from 24 to 39 months, $t(74) = 9.89$, $p = .000$. Average scores on the parent-report measures of temperament also showed significant improvement in children's Inhibitory Control, Attentional Focus, and Perceptual Sensitivity across this age span, $ts(59) = 4.76$, 6.07, and 9.66, respectively, $ps = .000$, whereas Impulsivity scores went down, $t(59) = -1.94$, $p = .057$.

Relation to Verbal Ability and Maternal Education Level

EF at Time 1 was significantly related to verbal ability assessed using the MacArthur CDI, $r(78) = .61$, $p = .000$. Similarly, EF at Time 2 was significantly related to scores on the PPVT-3, $r(81) = .41$, $p = .000$. Individual task correlations with verbal ability are shown in Tables 3 and 4. We included questions about parents' educational background at Time 2.[6] Maternal education was not related to EF Composite scores at Time 1, $r(81) = .04$, but it was significantly related to EF Composite scores at Time 2, $r(81) = .24$, $p = .033$. Note that maternal education level and children's verbal ability were not significantly correlated with each other: MacArthur CDI, $r(78) = .11$; PPVT-3, $r(81) = .08$.

Sex Differences

Consistent with previous findings of sex differences in EF (e.g., Carlson & Moses, 2001; Kochanska et al., 1996), we found that girls scored significantly higher than boys on the EF Composites at Time 1 and Time 2, $ts(79) = 2.83$ and 3.01, $ps = .006$ and .003, respectively. At 24 months, although girls received higher scores than boys on all five EF measures, there was not a significant multivariate effect of sex, $F(5, 75) = 1.75$, $p = .134$. This pattern was significant at 39 months, $F(8, 72) = 2.12$, $p = .045$.

[5] Children who completed less than half the test trials on a given task were considered refusals and so had missing data on that task. To handle missing data in the major analyses, we used a technique recommended by Graham, Cumsille, and Elek-Fisk (2003) for correlational and longitudinal data sets to replace missing data with ranked means across the nonmissing data for each participant on the items of interest. This method preserves individual differences.

[6] Father's education was not significantly related to any of the variables of interest in this study, and so only the results for mothers' education level are reported.

Table 1
Descriptive Statistics of Performance on All Measures at 24 and 39 Months

Measure	*M*	*SD*	Range	*N*
24 months				
MacArthur (language production total)	288.40	165.65	8–572	78
Reverse Categorization (proportion)	.42	.36	0–100	74
Multilocation Search (no. of perseverative errors, reverse scored)	3.26	1.06	0–4	74
Shape Stroop (average)	1.53	.41	0.67–2.00	74
Snack Delay (average wait in seconds)	7.33	4.71	0.90–12.50	80
Gift Delay (touch score)	3.24	1.67	1–5	79
Intentions (proportion)	.52	.38	0–1	72
Discrepant Desires (proportion)	.49	.41	0–1	74
Visual Perspective (average)	3.91	0.47	2.80–5.00	77
ISLQ (total)	23.91	23.06	0–89	74
TBAQ–R Inhibitory Control	4.04	0.80	2.29–6.07	77
TBAQ–R Attentional Focus	4.18	0.75	1.58–5.55	77
TBAQ–R Positive Anticipation	4.46	0.84	2.40–6.00	77
TBAQ–R Perceptual Sensitivity	4.04	1.03	1.69–6.46	77
39 months				
PPVT-3 (total)	56.68	15.62	10–102	81
Reverse Categorization (proportion)	.85	.30	.00–1.00	81
Grass/Snow (proportion)	.60	.34	.00–1.00	65
Snack Delay (composite)	0.00	0.88	−2.07–0.98	81
Bear/Dragon (composite)	0.00	0.93	−2.00–0.85	81
Hand Game (proportion)	.69	.39	.00–1.00	68
Tower Building (average)	18.85	15.19	−3.75–53.50	79
Whisper (average)	1.81	0.35	0.50–2.00	72
Gift Delay (composite)	0.00	0.93	−2.19–0.89	81
Pretend–Reality (proportion)	.68	.23	.33–1.00	81
Discrepant Desires (proportion)	.65	.31	.00–1.00	81
Visual PT Level 1 (proportion)	.88	.19	.17–1.00	81
Visual PT Level 2 (proportion)	.34	.32	.00–1.00	81
Appearance–Reality (proportion)	.36	.36	.00–1.00	81
Contents False Belief (proportion)	.21	.34	.00–1.00	81
Non-Mental-State Change (pass/fail)	.46	.50	0–1	79
CBQ Inhibitory Control (average)	4.70	0.96	2.33–6.67	62
CBQ Attentional Focus (average)	5.02	1.03	2.83–6.67	62
CBQ Impulsivity (average)	4.12	1.10	1.50–6.67	62
CBQ Perceptual Sensitivity (average)	5.41	1.02	3.00–7.00	62
Rules Questionnaire (average)	2.85	0.70	1.37–4.73	73

Note. MacArthur = MacArthur Communicative Development Inventory—Toddler Form; ISLQ = Internal States Language Questionnaire; TBAQ–R = Toddler Behavior Assessment Questionnaire—Revised; PPVT-3 = Peabody Picture Vocabulary Test—Third Edition; PT = Perspective Taking; CBQ = Children's Behavior Questionnaire—Short Form.

Table 2
Results of Principal-Components Analyses of Executive Function at 24 and 39 Months (N = 81)

24 months		39 months	
Task	Loading	Task	Loading
Snack Delay	.81	Bear/Dragon	.73
Shape Stroop	.64	Reverse Categorization	.70
Gift Delay	.61	Whisper	.65
Reverse Categorization	.55	Gift Delay	.63
Multilocation Search	.34	Hand Game	.62
		Tower Building	.60
		Grass/Snow	.54
		Snack Delay	.50

Note. Loadings are on the first unrotated principal component.

Follow-up univariate tests indicated girls outperformed boys on the Snack Delay, Gift Delay, Hand Game, and Bear/Dragon tasks (ps = .023, .005, .008, and .099, respectively). However, in a multivariate analysis of variance (MANOVA) with verbal ability covaried (PPVT-3), the multivariate effect of sex was no longer significant, $F(8, 71) = 1.58$, $p = .146$, which suggested that the difference is at least partially accounted for by girls' greater verbal ability.

Longitudinal Stability

Individual differences in EF Composite scores were moderately stable from 24 to 39 months, $r(81) = .51$, $p = .000$. This stability was independent of sex and verbal ability (PPVT-3), partial $r(77) = .38$, $p = .001$. Parent ratings of temperament character-

Table 3
Correlations Among Measures at Time 1 (24 months)

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Age	—	.17	.23*	.00	.00	.13	.16	.45**	.39**	.27*	.19†	−.08	.22†	.26*
2. Sex		—	.37**	.15	.15	.20†	.22*	.16	.17	.12	.04	.27*	.16	.38**
3. MacArthur			—	.11	.37**	.32**	.44**	.36**	.35**	.11	−.07	.24*	.42**	.83**
4. Maternal education				—	−.01	−.01	.03	.04	.10	−.11	−.06	.03	.04	.07
5. Reverse Categorization					—	−.02	.33**	.20†	.17	−.02	−.28*	.06	.11	.29*
6. Multilocation Search						—	.06	.19†	.17	.08	.10	.21†	.27*	.23*
7. Shape Stroop							—	.40**	.04	−.03	−.07	.09	.26*	.45**
8. Snack Delay								—	.43**	.02	.07	.08	.25*	.35**
9. Gift Delay (bow)									—	.08	.08	−.04	.06	.32**
10. Intentions										—	.44**	.20†	.19†	.03
11. Discrepant Desires											—	.07	.07	−.17
12. Visual Perspective												—	.25*	.15
13. Comprehension of Pretense													—	.43**
14. ISLQ														—

Note. MacArthur = MacArthur Communicative Development Inventory—Toddler Form; ISLQ = Internal States Language Questionnaire.
† $p < .10$. * $p < .05$. ** $p < .01$.

istics associated with the construct of self-control were also stable from 24 (TBAQ–R) to 39 months (CBQ): for Inhibitory Control, Attentional Focus, Impulsivity (with Positive Anticipation at Time 1), and Perceptual Sensitivity, *r*s(60) = .46, .40, .28, and .50, *p*s= .000, .002, .031, and .000, respectively.

Correspondence With Parent Report

Parent ratings of Attentional Focus (TBAQ–R) were positively related to children's EF Composite scores at 24 months, *r*(77) = .23, *p* = .043. Parent report of Inhibitory Control, Perceptual Sensitivity, and Positive Anticipation were unrelated to EF as this age. At Time 2, parent ratings of child temperament on three CBQ subscales were significantly related to contemporaneous EF Composite scores: for Inhibitory Control, Attentional Focus, and Perceptual Sensitivity, *r*s (62) = .26, .34, and .28, *p*s = .039, .007, and .029, respectively.

Next we examined whether temperament ratings at 24 months predicted EF Composite scores at 39 months. The subscale of Inhibitory Control on the TBAQ–R was significantly predictive, *r*(77) = .25, *p* = .031. No other subscale was significant. In the converse analysis, 24-month EF Composite scores were unrelated to later temperament ratings on the CBQ; *r*s(62) ranged from −.19 to .20, *ns*. Finally, average scores on the Rules Questionnaire (administered at 39 months) were not significantly related to children's EF Composite scores at 24 or 39 months, *r*s(73) = −.12 and −.19, respectively, *ns*. However, the fact that the correlations were negative could be due to household rules being considered more important to enforce with children who have difficulty with self-control.

Theory of Mind

Descriptive statistics for each ToM task at Times 1 and 2 are shown in Table 1.

Data Reduction

We conducted principal-components analyses to inform data reduction of the ToM measures at each time point. As shown in Table 5, all four ToM measures at 24 months had item loadings ≥ .52 on the FUPC. Therefore, we standardized and averaged scores across tasks to form composite ToM scores at Time 1. The Intention and Discrepant Desires tasks were particularly highly correlated (see Table 3). This relation remained significant after we controlled for age, sex, and verbal ability, partial *r*(66) = .36, *p* = .002. However, the internal consistency of the four measures was modest (Cronbach's α = .50), and so we also examined the four individual tasks separately in major analyses.

At 39 months, children were given six ToM tasks. Of these, two had low item loadings on the FUPC: Discrepant Desires and Visual Perspective Taking Level 2. The other four tasks (Pretend–Reality, Appearance–Reality, Visual Perspective Taking Level 1, and False Belief) had loadings ≥ .48 (see Table 5). We aggregated the standardized scores on these four measures, thus creating a ToM Composite score for each participant at Time 2. Coherence was again modest (α = .50), and thus individual tasks were considered as well. Bivariate correlations among the individual ToM measures are shown in Table 4. Appearance–Reality and Pretend-Reality scores were the most highly correlated and remained so independent of age, sex, and verbal ability, *r*(76) = .30, *p* = .007.

Relation to Verbal Ability and Maternal Education Level

ToM Composite scores at Time 1 and Time 2 were significantly related to verbal ability, *r*s(78 and 81) = .25 and .44, *p*s = .027 and .000, respectively. At 24 months, Visual Perspective and Comprehension of Pretense were both significantly related to language production (MacArthur CDI scores), as shown in Table 3. At 39 months, Pretend–Reality and Visual Perspective Taking Level 1 were significantly related to verbal ability (PPVT-3), and False Belief almost reached significance (see Table 4). Theory of mind was not related to maternal education level at either time point. This was true for the aggregates as well as the individual measures.

Sex Differences and Development

There was some evidence of sex differences favoring girls on the ToM measures. At Time 1, girls had significantly higher ToM

Table 4
Correlations Among Measures at Time 2 (39 months)

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Age	—	.12	.00	−.21†	.00	−.01	.09	−.03	−.14	−.01	.17	.02	.09	−.18	−.07	−.04	.10	−.06	−.18
2. Sex		—	.24*	.15	.17	.03	.25*	.18†	.29**	.06	.16	.31**	.29**	.03	−.04	−.06	.13	−.04	−.01
3. PPVT-3			—	.08	.11	.23*	.14	.39**	.27*	.09	.17	.43**	.39**	.05	.31**	−.07	.16	.21†	.20†
4. Maternal education				—	.24*	.08	.14	.15	.19†	.08	.23*	.07	−.04	.06	.10	.14	−.10	−.09	.16
5. Reverse Categ.					—	.31**	.38**	.37**	.30**	.30**	.45**	.33**	.37**	.03	.27*	.05	.10	.02	.13
6. Grass/Snow						—	.19†	.24*	.30**	.32**	.31**	.13	.32**	−.09	.24*	−.05	.15	.21†	.33**
7. Snack Delay (choice)							—	.20†	.17	.27*	.20†	.24*	.29**	.24*	.15	−.11	.31**	−.02	.06
8. Bear/Dragon								—	.52**	.33**	.40**	.43**	.54**	.06	.32**	−.10	.22†	.06	.36**
9. Hand Game									—	.24*	.26*	.28*	.23*	−.05	.32**	−.02	.10	.18	.27*
10. Tower Building										—	.21†	.34**	.27*	.04	.28*	−.03	.18	−.03	.15
11. Whisper											—	.35**	.37**	.04	.21†	−.14	.02	.03	.26*
12. Gift Delay												—	.32**	.10	.17	−.13	.22†	.02	.14
13. Pretend–Reality													—	−.02	.27*	−.16	.35**	.11	.26*
14. Discrepant Desires														—	.13	.08	−.03	.11	.01
15. Visual PT Level 1															—	.09	.01	.08	.17
16. Visual PT Level 2																—	.06	.24*	−.10
17. Appearance–Reality																	—	.25*	.11
18. False Belief																		—	.25*
19. Non-Mental-State Change																			—

Note. PPVT-3 = Peabody Picture Vocabulary Test—Third Edition; Categ. = Categorization; PT = Perspective Taking.
† $p < .10$. * $p < .05$. ** $p < .01$.

Composite scores than boys, $t(79) = 2.32$, $p = .023$. This difference was nonsignificant, however, with verbal ability covaried. There were no other significant sex differences in ToM at 24 months. At Time 2, there was no sex difference on the ToM Composite. The only significant sex difference at this age was on Pretend–Reality, $F(1, 80) = 7.43$, $p = .008$. However, this result fell below significance after we controlled for age and PPVT-3 scores. These results suggest, similarly to the EF results, that the observed sex differences in ToM were at least partially mediated by verbal ability. There was significant improvement over time on Discrepant Desires, the only directly comparable ToM task (see Table 1 for descriptives), $t(73) = 2.76$, $p = .007$.

Longitudinal Stability

Next, we examined the relations between ToM measures at Time 1 and Time 2 to assess whether individual differences in "early ToM" at age 2 were predictive of later ToM performance. The overall composites were not significantly related, $r(81) = .11$. However, the ToM Composite, Comprehension of Pretense, and Visual Perspective at Time 1 each significantly predicted Pretend–Reality scores at Time 2, rs(81) = .24, .22, and .26, ps = .029, .05, and .018, respectively. Results for the ToM Composite at Time 1 and Comprehension of Pretense were marginally significant after we controlled for PPVT-3 scores, rs(78) = .21 and .19, ps = .062 and .098, respectively. The Visual Perspective task fell below significance in this analysis. No ToM measures at Time 2 aside from Pretend–Reality were significantly predicted by Time 1 ToM behavioral measures.

Correspondence With Parent Report

Parents completed the ISLQ at Time 1. This questionnaire assesses the number of internal-state words in children's productive vocabularies. ISLQ scores at 24 months were significantly related to concurrent performance on the Comprehension of Pretense task (only), as shown in Table 3. This relation remained significant after we controlled for sex and PPVT-3 scores, $r(70) = .39$, $p = .001$. The ISLQ also predicted ToM Composite and Pretend–Reality scores at 39 months, rs(74) = .31 and .41, ps = .007 and .000, respectively. In partial correlations controlling for sex and PPVT-3, the relation between the ISLQ and Pretend–Reality scores remained significant, $r(70) = .25$, $p = .035$. The

Table 5
Results of Principal-Components Analyses of Theory of Mind at 24 and 39 Months (N = 81)

24 Months		39 Months	
Task	Loading	Task	Loading
Intentions	.79	Appearance–Reality	.68
Discrepant Desires	.67	Pretend–Reality	.67
Visual Perspective	.53	False Belief	.60
Comprehension of Pretense	.52	Visual PT Level 1	.48
		Visual PT Level 2	.23
		Discrepant Desires	.18

Note. Loadings are on the first unrotated principal component. PT = Perspective Taking.

ISLQ–ToM Composite relation was no longer significant (r = .18). Thus, the ISLQ was a useful parent-report measure of early ToM that was significantly related to children's later understanding of the mental state of pretense, in particular, and this was independent of general verbal ability. We therefore included the ISLQ in the major analyses of the relation between EF and ToM that follow.

Relation Between Executive Function and Theory of Mind

Next we examined the critical questions of when EF–ToM relations emerge, whether individual differences in EF would predict children's later ToM, whether early ToM abilities would predict later EF, and whether there was evidence of a reciprocal relation between the two constructs in development.

Time 1

At 24 months, EF Composite scores were positively but nonsignificantly related to the ToM Composite (see Table 6). Of the individual ToM measures, Comprehension of Pretense was significantly related to the EF Composite and was marginally significant in a partial correlation analysis controlling for age, sex, and verbal ability. Individual EF and ToM task correlations are shown in Table 3. These analyses revealed that Comprehension of Pretense was significantly related to several executive tasks, including Multilocation Search, Shape Stroop, and Snack Delay. When we controlled for age, sex, and verbal ability, the relation remained significant for Multilocation Search, $r(76) = .23$, $p = .046$, almost reached significance for Shape Stroop, $r(76) = .20$, $p = .078$, but was no longer significant for Snack Delay ($r = .15$, *ns*). In contrast to the behavioral measures, parent report of internal-state language was significantly related to children's executive functioning (EF Composite), and this relation held up when age, sex, and verbal ability were partialled out (see Table 6). Table 3 shows that the ISLQ was significantly related to all five EF tasks at Time 1. After we controlled for age, sex, and verbal ability, the relation with Shape Stroop remained significant, $r(69) = .36$, $p = .002$, whereas relations with Reverse Categorization and Snack Delay were not quite significant, rs(69) = .20 and .20, ps = .094 and .091, respectively, and relations with Multilocation Search and Gift Delay were nonsignificant (rs = .12 and .14, respectively). Of the four temperament scales we examined using the TBAQ–R (Positive Anticipation, Attentional Focus, Inhibitory Control, and Perceptual Sensitivity), Attentional Focus and Perceptual Sensitivity were significantly related to ISLQ scores, rs(72) = .35 and .24, ps = .002 and .043, respectively, and remained significant after we controlled for age, sex, and verbal ability, rs(67) = .36 and .23, ps = .002 and .055, respectively. No other ToM measures at Time 1 were related to temperament.

To summarize, at 24 months there was limited evidence for a relation between EF and ToM. One out of four behavioral measures of children's awareness of mental states (Comprehension of Pretense) and the parent-report measure of children's use of words to represent internal, nonobservable states were significantly related to some of the behavioral and parent-report measures of executive functioning. The parent-report ISLQ demonstrated the strongest concurrent relation to EF in toddlers, independent of effects due to age, sex, and verbal ability.

Time 2

At 39 months, however, the overall pattern of results was stronger. EF was significantly related to ToM Composite scores even after we controlled for age, sex, and PPVT-3 (see Table 6). As expected, children's performance on the Non-Mental-State task was significantly related to both EF and ToM at Time 2, rs(79) = .33 and .32, ps = .003 and .004, respectively. Even when we controlled for scores on this task in addition to the other controls, however, EF and ToM remained significantly related, partial $r(73) = .37$, $p = .001$, indicating that the relation was not accounted for by simple memory or inhibition problems. There were several significant individual task correlations as well, which are displayed in Table 4. Pretend–Reality was significantly related to all eight EF tasks. Appearance–Reality was significantly related to Snack Delay and marginally related to Gift Delay and Bear/Dragon. False Belief was significantly related to Grass/Snow. Finally, Visual Perspective Taking Level 1 was significantly related to children's performance on Reverse Categorization, Grass/Snow, Hand Game, Bear/Dragon, and Tower and was marginally related to their performance on Whisper. None of the four CBQ

Table 6
Correlations Between Executive Function Composite Scores and Theory of Mind (ToM) Measures at Time 1 and Time 2

Time 1 (24 months)			Time 2 (39 months)		
ToM measure	r	Partial r	ToM measure	r	Partial r
ToM Composite	.17	.01	ToM Composite	.50**	.41**
Intentions	.05	−.12	Pretend–Reality	.55**	.43**
Discrepant Desires	−.02	−.07	Appearance–Reality	.27*	.21†
Visual Perspective	.12	.01	False Belief	.14	.08
Comprehension of Pretense	.31**	.21†	Visual PT Level 1	.34**	.29**
ISLQ	.55**	.38**	Visual PT Level 2	−.08	−.04
			Discrepant Desires	.04	.01

Note. Partial correlations control for age, sex, and verbal ability. Alpha level = .05. ISLQ = Internal States Language Questionnaire; PT = Perspective Taking.
† $p < .10$. * $p < .05$. ** $p < .01$.

temperament scales we examined was significantly related to ToM Composite scores at Time 2.

Prediction From Time 1 to Time 2

Next, we addressed the question of whether early self-control predicts individual differences in ToM, and the converse, whether early mental representation skills predict later EF performance. First we examined the correlation between EF at Time 1 and ToM at Time 2 (composites). This relation was significant, $r(81) = .42$, $p = .000$. Importantly, the relation remained significant even after we controlled for EF scores at Time 2, $r(78) = .26$, $p = .018$, indicating that the longitudinal relation between EF and ToM was not entirely accounted for by high stability in the EF construct itself. Next, we carried out a hierarchical multiple regression analysis (using the enter method) in which the basic controls were entered in the first block, the measures of ToM at Time 1 were entered in the second block, and the major predictor of interest, EF at Time 1, was entered in the final block. The results are provided in Table 7. This analysis revealed that after numerous controls were accounted for, EF performance at 24 months significantly predicted ToM Composite scores at 39 months. Verbal ability (PPVT-3) also made a unique contribution to the variance in ToM in this analysis. The final model was significant, $F(7, 73) = 3.77$, $p = .002$, $R^2 = .29$, as was the change in R^2 when the effects of EF at Time 1 were tested after all other relevant variables were accounted for. It was possible that including the ToM Composite at Time 1 was obscuring these results because, unlike the ISLQ, it was not significantly related to the criterion (ToM Composite at Time 2). However, the results for EF at Time 1 were unchanged when ISLQ scores were entered alone in the second block.

Recall that the early ToM measures best predicted Pretend–Reality scores of all the Time 2 ToM tasks. In a separate regression analysis in which Pretend–Reality scores were used as the criterion instead of the ToM Composite, verbal ability was again significant in the first block, $\beta = .38$, $p = .001$, $\triangle R^2 = .22$, $p = .002$. In the second block, the parent-report (ISLQ) measure of ToM was significant ($\beta = .24$, $p = .042$) in addition to verbal ability ($\beta = .29$, $p = .011$), $\triangle R^2 = .08$, $p = .031$. In the final block, EF at Time 1 was again a significant predictor, $\beta = .35$, $p = .005$, $\triangle R^2 = .08$, $p = .005$. Verbal ability, but not the Time 1 ToM measures, remained significant, $\beta = .24$, $p = .03$. The final model was significant, $F(6, 73) = 5.72$, $p = .000$, $R^2 = .39$. In parallel regression analyses examining each of the other three ToM tasks included in the composite at Time 2, we found that EF at Time 1 (entered last) was a marginal predictor of Appearance–Reality ($\beta = .24$, $p < .10$), but not False Belief ($\beta = .08$) or Visual Perspective Taking Level 1 ($\beta = .01$). None of the TBAQ–R temperament subscales of interest significantly predicted later ToM performance.

Table 7
Summary of Hierarchical Regression Analysis for Variables Predicting 39-Month Theory of Mind (ToM) Composite Scores (Final Model)

Variable	B	$SE\ B$	β	ΔR^2
Block 1				.21**
Age	2.17	2.85	.08	
Sex	−1.84	5.54	−.04	
PPVT-3	.48	.18	.32**	
Maternal education	−2.69	3.22	−.09	
Block 2				.03
ToM Time 1	.00	.11	.00	
ISLQ Time 1	.00	.14	.06	
Block 3				.05*
EF Time 1	.28	.13	.28*	

Note. $N = 73$. Alpha level = .05. PPVT-3 = Peabody Picture Vocabulary Test—Third Edition; EF = Executive Function Composite; ISLQ = Internal States Language Questionnaire.
* $p < .05$. ** $p < .01$.

To summarize, executive control at 24 months and verbal ability each made significant and unique contributions to individual differences in ToM at 39 months (ToM Composite at Time 2). The results for EF at Time 1 held up even when we used Pretend–Reality scores as the criterion. This test was more stringent because ToM at Time 1 was significantly related to Pretend–Reality at Time 2 (i.e., evidence of some stability in ToM) when we controlled for verbal ability. In this analysis, once we added EF at Time 1 in the final step, it was highly significant over and above the contributions of ToM at Time 1 and verbal ability.

The preceding results suggested there was evidence of longitudinal prediction from early individual differences in EF to children's later proficiency in ToM. We then tested the reciprocal relation, that is, whether early ToM predicted later EF. First, we conducted correlation analyses and found that ToM at Time 1 (the behavioral battery) was unrelated to EF at Time 2, $r(81) = -.01$. Of the four individual ToM tasks, only Visual Perspective was significantly predictive of EF, $r(81) = .27$, $p = .014$, but this was nonsignificant after we controlled for age, sex, and verbal ability ($p = .14$).

In contrast, the ISLQ (parent report) was significantly related to later EF scores, $r(74) = .47$, $p = .000$, and remained so after we partialed out ToM at Time 2, $r(71) = .38$, $p = .001$ (i.e., controlling for longitudinal stability in ToM itself). Next we carried out a multiple regression that included the basic controls in the first block, EF at Time 1 in the second block, and ISLQ scores in the final block. These results are shown in Table 8. We found that verbal ability and EF performance at Time 1 were significant predictors of EF at Time 2, and maternal education was a marginal predictor, but ISLQ scores were not significant predictors over and above these other variables. The final model was significant, $F(6, 73) = 7.48$, $p = .000$, $R^2 = .40$. These analyses suggested that although children's use of internal-state words at 24 months was associated with later executive control in correlation analyses, this relation did not hold up independently of EF at Time 1 and receptive vocabulary. The ToM measures at 24 months also did not significantly predict later temperament ratings on the CBQ subscales associated with self-control. Hence, the 24-month ToM Composite, individual measures, and the parent-report measures all failed to provide clear evidence of a reciprocal relation.

Task Specificity

Last, having found a significant relation between EF and ToM at Time 2 as well as prediction of ToM from EF at Time 1, we turned to the question of which individual EF tasks were most strongly related to ToM. As illustrated in Table 9, the strongest

Table 8
Summary of Hierarchical Regression Analysis for Variables Predicting 39-Month Executive Function Composite Scores (Final Model)

Variable	*B*	*SE B*	β	ΔR^2
Block 1				.26**
Age	.88	2.58	.03	
Sex	2.16	4.87	.05	
PPVT-3	.34	.15	.23*	
Maternal education	5.08	2.83	.18†	
Block 2				.12**
EF Time 1	.31	.11	.31**	
Block 3				.02
ISLQ Time 1	.18	.12	.18	

Note. $N = 73$. Alpha level = .05. PPVT-3 = Peabody Picture Vocabulary Test—Third Edition; EF = Executive Function Composite; ISLQ = Internal States Language Questionnaire.
† $p < .10$. * $p < .05$. ** $p < .01$.

predictors of ToM (Composite scores) from EF at 24 months were Reverse Categorization and Snack Delay (after we controlled for age, sex, and verbal ability). To determine whether one or more tasks was uniquely related to ToM, we conducted a hierarchical multiple regression similar to that for the EF Composite at Time 1 predicting ToM (see Table 7) except that all five EF tasks from Time 1 were entered in the final block. Snack Delay and Reverse Categorization were significant in this analysis (βs = .33 and .24, *p*s = .011 and .050, respectively). Verbal ability (PPVT-3) was also significant, $\beta = .31$, $p = .009$. The final model was significant, $F(11, 73) = 3.03$, $p = .003$, $R^2 = .35$. The EF measures at Time 1 accounted for 13% of the variance in ToM at Time 2 in this analysis ($p = .04$).

Turning to the 39-month data, seven of the EF tasks were significantly related to ToM, and each of these remained significant or almost significant in partial correlations, as shown in Table 9. Only one task—Gift Delay—was clearly unrelated to ToM Composite scores after the relevant controls were added. When we entered all eight EF tasks at Time 2 in a hierarchical multiple regression analysis, after controlling for age, sex, PPVT-3, and maternal education, we found that two individual EF tasks were uniquely related to concurrent ToM performance: Bear/Dragon and Snack Delay, βs = .28 and .26, *p*s = .034 and .02, respectively. Verbal ability (PPVT-3) also was significant, $\beta = .29$, $p = .012$. The final model was significant, $F(12, 80) = 3.63$, $p = .000$, $R^2 = .39$. The EF measures at Time 2 (entered last) accounted for 18% of the variance in ToM ($p = .018$).

Hence, at both 24 and 39 months, there was evidence for two types of EF task sharing unique variance with ToM (Time 2). Conflict tasks call for the suppression of a dominant response and the initiation of a subdominant response (as in Reverse Categorization) or an alternation between performing and suppressing dominant responses (as in Bear/Dragon). In contrast, delay tasks require children to withhold a dominant response over a temporal delay (as in Snack Delay). We therefore examined the relative contributions of EF tasks at Time 2 that were designed to measure conflict and delay, given that this distinction was important in previous research on EF–ToM relations (e.g., Carlson & Moses, 2001; Carlson et al., 2002; Hala et al., 2003). We computed separate scores for the five conceptually derived conflict tasks (Hand Game, Grass/Snow, Bear/Dragon, Reverse Categorization, and Whisper) and the three delay tasks (Snack Delay, Tower, and Gift Delay). Both task batteries were significantly related to the ToM Composite, both *r*s(81) = .42, *p*s = .000. These relations remained significant after we controlled for age, sex, and verbal ability, partial *r*s(76) = .32 (conflict tasks) and .33 (delay tasks), *p*s = .004 and .003, respectively. Note that the conflict and delay batteries were also significantly correlated with one another, $r(81) = .54$, $p = .000$. Next, we directly compared these two aspects of EF in a hierarchical multiple regression, entering the usual controls in the first block, delay EF in the second block (which was significant, $\beta = .33$, $p = .002$), and conflict EF in the final block. The results of the final model are shown in Table 10. The conflict battery was significantly related to ToM over and above delay EF and the controls, whereas the delay battery was not significant over and above the conflict battery. The final model was significant, $F(6, 80) = 6.38$, $p = .000$, $R^2 = .34$. Thus, although there was evidence of certain individual conflict and delay tasks explaining unique variance in ToM at age 3, in keeping with previous research on this topic, only the conflict battery remained significant when these variables were pitted against one another.

Table 9
Individual Executive Function Task Correlations With 39-Month Theory of Mind

24 months			39 months		
Task	*r*	Partial *r*	Task	*r*	Partial *r*
Snack Delay	.39**	.33**	Bear/Dragon	.46**	.35**
Reverse Categorization	.29**	.21†	Snack Delay	.36**	.33**
Multilocation Search	.22*	.18	Grass/Snow	.34**	.28*
Shape Stroop	.20†	.11	Reverse Categorization	.29**	.27*
Gift Delay	.16	.02	Hand Game	.29**	.22†
			Tower Building	.25*	.24*
			Whisper	.25*	.19†
			Gift Delay	.31**	.15

Note. Partial correlations control for age (Time 2), sex, and verbal ability. Alpha level = .05.
† $p < .10$. * $p < .05$. ** $p < .01$.

Table 10
Summary of Hierarchical Regression Analysis for Conflict and Delay Aspects of Executive Function (EF) Predicting 39-Month Theory of Mind Composite Scores (Final Model)

Variable	*B*	*SE B*	β	ΔR^2
Block 1				.21**
Age	1.23	2.54	.05	
Sex	−1.51	4.76	−.03	
PPVT-3	.46	.16	.31**	
Maternal education	−5.24	2.99	−.18	
Block 2				.09**
Delay EF	7.10	3.75	.22	
Block 3				.04*
Conflict EF	8.51	4.02	.25*	

Note. $N = 80$. Alpha level = .05. PPVT-3 = Peabody Picture Vocabulary Test—Third Edition.
* $p < .05$. ** $p < .01$.

Discussion

In this longitudinal study we provide three main contributions to the research on children's EF and ToM. First, we examined the coherence and stability of EF and ToM using both new and established behavioral measures. Coherence of the measures was modest for both EF and ToM at 24 months. At 39 months, EF but not ToM had high internal consistency. Individual differences in EF were stable over time, whereas for ToM the stability was limited to individual tasks. Second, we found that a robust relation between individual differences in EF and ToM did not emerge until the 39-month assessment. Third, EF performance in children as young as 24 months significantly predicted performance on ToM tasks 1 year later, even after we controlled for other variables including verbal ability and maternal education. In contrast, there was limited evidence of a reciprocal relation in which internal-state language predicted later EF, but this did not hold up over the control variables. These findings raise the exciting possibility of identifying developmental precursors of children's mental-state awareness (and potential impairments therein) from a much earlier age than has been previously explored.[7]

The stability of individual differences in EF performance is consistent with Kochanska et al.'s (2000) findings in a longitudinal study of effortful control from 22 to 33 months of age (see also Vaughn, Kopp, & Krakow, 1984). Our study also benefited from their having developed and validated several of the EF measures we used. Given the high stability of EF in early childhood, it would be desirable to establish age norms on standardized tasks for future research on EF development and its social and cognitive correlates and consequences. Currently no norms exist for children below age 3 (e.g., the NEPSY; Korkman, Kirk, & Kemp, 1998).

In contrast, there was limited evidence for coherence and stability of individual differences in ToM in this study. Internal consistency was modest at both time points, and so we presented results for individual tasks as well as the composites. ToM coherence at 39 months was lower than what has been found in studies of older preschoolers. For example, Carlson and Moses (2001) reported an alpha of .76 for four ToM measures (two types of false belief, deception, and appearance–reality) in a sample of 107 children with a mean age of 3 years 11 months. It is interesting to note that with few exceptions in the ToM literature, investigators routinely compute aggregate ToM scores simply by summing across the various ToM tasks. "Standard" ToM tasks such as the appearance–reality and false belief tasks are assumed to reflect a common conceptual core. Indeed, several studies have confirmed this empirically, showing that various subsets of ToM tasks are significantly intercorrelated, even when age and, in some studies, verbal ability are controlled (e.g., Carlson & Moses, 2001; Flavell, Green, & Flavell, 1989; Frye et al., 1995; Gopnik & Astington, 1988; Moore, Pure, & Furrow, 1990; Taylor & Carlson, 1997; Wimmer & Hartl, 1991). Poorer coherence has been reported in other studies, however, in which not all standard ToM tasks were significantly correlated (e.g., Carlson et al., 2002; Hala et al., 2003; Hughes, 1998a). More important, in longitudinal studies with children slightly older than those in our study, the association between early and late ToM scores was nonsignificant after age at Time 1 was controlled for (Hughes, 1998b), or there were periods of regression in performance on ToM tasks (Flynn, O'Malley, & Wood, 2004).

Some researchers (e.g., Wellman, 1990) have suggested that an early understanding of goals and intentions grows into the belief–desire reasoning characteristic of ToM in preschoolers. Our findings did not support this hypothesis, but this could be due to measurement issues, inasmuch as the children in our study were older than the children in the original intention and desire studies, which both tested 18-month-olds (Meltzoff, 1995; Repacholi & Gopnik, 1997). We modified the original procedures to afford greater variability in older children, but it is possible that these tests are more sensitive at the younger ages for which they were originally designed. An alternative interpretation is that an understanding of others' goals and intentions does not directly feed into children's mental representational competence. Recall that even at 39 months, the modified Discrepant Desires task was not related to the other ToM tasks, despite adequate variability. These results are consistent with those of Ziv and Frye (2003), who reported that 3-year-olds relied on different cues to infer desire than to infer belief. Desire was inferred from agents' links to objects or from objects' common desirability, whereas belief inferences corresponded to the current state of affairs. Furthermore, although 3-year-olds were accurate in desire inferences, they did not use them for inferring belief within the same person or situation. Tager-Flusberg and Sullivan (2000) also found that the "social-perceptual" (inferring what someone is attending to or wants) and "social-cognitive" (understanding belief) aspects of ToM were dissociable in children with Williams syndrome. Interestingly, in recent neuroimaging research with normal adults, Saxe, Carey, and Kanwisher (2004) reported that stories involving the interpretation of actors' goals and intentions and stories involving false beliefs activated adjacent but distinct areas of bilateral temporo-parietal cortex. It is possible that in development, these skills may be part of a shared domain (i.e., ToM) but have proximal rather than identical neural substrates.

[7] We conducted cross-validation analyses to assess the replicability of our results. Two random subsamples were compared (*N*s = 41 and 40, roughly equal numbers of boys and girls). These analyses revealed that (a) the two subgroups had highly consistent results, and (b) all major analyses produced results virtually identical to those of the sample as a whole, although in some cases they fell below statistical significance because of reduced power.

In contrast to the behavioral measures, we found that the ISLQ (Bretherton & Beeghly, 1982) at 24 months significantly predicted later ToM Composite scores. This finding is consistent with previous research showing that early exposure to mental- and emotional-state talk in the home contributes to the development of ToM (e.g., J. Dunn, et al., 1991; Hughes & Dunn, 1998; Meins et al., 2002; Ruffman et al., 2002). The ISLQ was most strongly predictive of performance on the Pretend–Reality task (independent of general language proficiency), which may index the kind of mental–physical distinction that young 3-year-olds are most attuned to or motivated to understand (e.g., frequently asking parents to confirm that something is "just pretend").

A second contribution of our research is that a strong EF–ToM relation was not apparent until the 39-month assessment. The results for the 3-year-olds are highly consistent with a growing body of research showing a relation between EF and ToM in the preschool period (e.g., Carlson & Moses, 2001; Frye et al., 1995; Hughes, 1998a; Perner & Lang, 1999; Perner et al., 2002). The present study extends evidence for this link to a sample consisting exclusively of young 3-year-olds and used several EF and ToM measures different from those used in the preschool studies. We can conclude that the close association between executive processes and ToM in preschoolers is indeed a robust one.

Our inability to detect concurrent relations between EF and ToM in the behavioral measures at 24 months might circle back to the challenge we described in the introduction of measuring ToM in toddler-age children. A lack of coherence in the ToM tasks could indicate there was insufficient sensitivity of measurement to detect relations with EF. An alternative explanation is that there is genuinely no relation between EF skills and the early-developing intention–desire aspects of ToM. Moses and Carlson (in press) proposed a theoretical account in which EF *interacts* with conceptual content, such that it will relate strongly to children's understanding of mental states such as belief that carry a deep-seated reality bias; the very point of belief is to faithfully represent the true state of affairs (Carlson et al., 1998; Moses, 1993). In contrast, in this account, EF would be only weakly associated with mental states such as intention and desire, in which there is less pressure to align with reality. Moses and Carlson (in press) described recent evidence from preschoolers that is consistent with this hypothesis. Hence, EF–ToM relations might not emerge until children begin to wrestle with mental states that contradict their expectations regarding reality. Interestingly, the one exception to the low EF–ToM correlations at 24 months was the Comprehension of Pretense task (Harris & Kavanaugh, 1993). This task was significantly related to concurrent EF Composite scores as well as three individual EF tasks. In the partial correlations controlling for verbal ability, the relation with Multilocation Search (Zelazo et al., 1998) remained significant. Comprehension of Pretense also predicted Pretend–Reality scores 1 year later. As we alluded to earlier, young preschoolers might be most prone to grappling with the distinction between pretense and reality prior to belief. At 24 months (and to some extent at 39 months) children are beginning to come to grips with the "point" of pretense being purposefully *anti*-reality. Our data suggest this distinction might be the best place to look first for executive functioning as a constraint on early-developing concepts about the mind.

The parent-report ISLQ painted a different picture from the one shown by the laboratory tasks. Although not a ToM task proper, this index of children's internal-state word usage was significantly related to EF performance at both 24 and 39 months when verbal ability was controlled. This finding is exciting because it suggests that similar processes (language, mental-state talk) that have been shown to promote ToM are also contributing to the development of executive function. Several researchers have posited that verbal regulation of behavior in the form of self-speech is an essential ingredient of self-control (Barkley, 1997; Berk, 1992; Carlson, 2003; Vygotsky, 1962; Zelazo, 1999).

The third and most significant contribution of our research concerns the longitudinal relations between EF and ToM. On this point, our results are largely consistent with those of Hughes (1998b). She found that performance on a detour reaching task, thought to assess the inhibition of a maladaptive prepotent response, was a unique longitudinal predictor of ToM (assessed using false belief and perspective-taking tasks) from 3 years 11 months to 5 years 0 months of age. Our research extends Hughes's (1998b) findings in an important way by showing that a similar pattern of results was observed in a much younger sample of children. Similarly, Flynn et al. (2004) conducted a microgenetic study of inhibition skills and false belief task performance in a sample of 21 children every 4 weeks (across six phases) beginning when the children were about 3.5 years old. They found that mastery of inhibitory control (as measured by Luria's Hand Game and Lights Task; Luria et al., 1964) developmentally preceded successful performance on false belief. Our results with a younger age group and a longer lag between assessments are broadly consistent with their findings.

Similarly to Hughes's (1998b) study, our study did not support a reciprocal relation between EF and ToM over the course of development. Specifically, we found that although the ISLQ score at Time 1 predicted EF at Time 2 when verbal ability was controlled, this relation did not hold up over and above EF at Time 1. In other words, we could not rule out the possibility that the concurrent relation between mental-state word usage and executive control at 24 months and/or the stability of the EF construct itself could account for the longitudinal result. In contrast, EF at Time 1 predicted ToM at Time 2, over and above ISLQ scores in addition to the other controls. The EF $\rightarrow$ ToM prediction held up even when we controlled for EF at Time 2, that is, controlled for the strong concurrent relation between the two constructs. The present study therefore did not fully support the hypothesis that the ability to meta-represent, as exhibited in ToM, is necessary for children to successfully inhibit a prepotent response and execute a desired action on tests of executive control (Perner, 1998; Perner & Lang, 2000). Instead, we observed an asymmetry in which children's relative standing on EF measures at age 2 predicted their relative standing on ToM at age 3 better than the reverse.

At Time 2, we saw a familiar pattern in which the conflict EF measures were significantly related to ToM even after accounting for the relation between ToM and delay EF. Conflict measures included tasks such as Bear/Dragon, Reverse Categorization, and Grass/Snow, in which successful performance requires one to inhibit a prepotent response and adhere to a set of rules. Carlson and Moses (2001) reported similar findings in 3- and 4-year-olds. They proposed that the conflict tasks combine inhibitory control and working memory demands, whereas delay tasks (e.g., Gift Delay) have relatively attenuated working memory demands (e.g., "don't peek"). ToM performance is consistently more strongly related to EF measures involving both inhibition and working memory (Carlson et al., 2002; Hala et al., 2003). Diamond, Kirkham, and Amso (2002) showed experimentally that the Day/Night Stroop-like task, which is analogous in structure to the Grass/Snow task, depends critically on a combination of inhibition and working memory. Our research suggests the same may be true of most ToM tasks.

This pattern of findings is relevant to the cognitive complexity and control interpretation of EF–ToM relations, in which ToM tasks such as False Belief and Appearance–Reality are considered a type of executive task having a complex rule structure (Frye, 2000; Frye et al., 1998; Zelazo, Müller, Frye, & Marcovitch, 2003). According to this account, executive tasks that also have an embedded rule structure, such as conflict tasks, will be most strongly related to ToM. The present results and those of previous studies are consistent with this analysis (e.g., Carlson et al., 2002; Hala et al., 2003). Some evidence from the present study did not fit neatly with a conflict–delay distinction, however. At 24 months, one conflict task (Reverse Categorization) and one task conceptualized as delay of gratification (Snack Delay) were significant in predicting ToM 1 year later when all EF tasks were included in a regression analysis. Similarly, both a conflict task (Bear/Dragon) and a delay task (Snack Delay) explained unique variance in ToM at Time 2. We speculate that appetitive self-control might play a greater role earlier in development, when children's working memory capacity is still meager. Recall that at 39 months the Snack Delay procedure followed that of Mischel et al. (1989), in which children had a choice to wait to receive the larger reward. It is possible that this task could be said to have a more complex rule structure according to Zelazo et al. (2003) or to require more working memory than simple delay tasks. However, the fact that we found that at least *some* delay tasks were significantly related to ToM over and above individual conflict tasks goes against the hypothesis that the ability to reason with embedded rules is the key development linking EF and ToM.

We propose that, on balance, the cumulative evidence thus far indicates that EF is a potent, although not exclusive, contributor to children's developing understanding of mind. A limitation of this study is that the internal consistency of the behavioral ToM tasks was low at both time points, as was longitudinal stability in this construct. Although we attempted to select measures that were conceptually related, a dearth of research on ToM with 24- and 39-month-olds meant we were in uncharted territory. We addressed this shortcoming by examining the aggregated and individual ToM measures separately in major analyses, which revealed largely consistent results. In contrast to the behavioral measures, parent report of children's internal-state language at 24 months was moderately stable with later ToM, was related to EF concurrently, and predicted EF at Time 2, although not over the controls. It remains possible that we have not identified a ToM task or set of tasks that would uncover a significant relation to EF at 2 years of age and that would uniquely predict children's later executive functioning. Although our results cannot definitively rule out Perner's (1998) claim that EF is functionally dependent on ToM, they are certainly consistent with Russell's (1996) proposal that controlling attention and inhibiting prepotent responses serve children's growing capacity to reflect on the mental states of self and others (i.e., that ToM may be functionally dependent on EF). Our research indicates that individual differences in EF matter for children's ToM development and that a predictive relation can be traced from at least 24 months of age.

Awaiting further study is the mechanism through which EF affects ToM development. It could have a direct influence, such that a well-functioning executive system helps concepts of mind get off the ground (Russell, 1996). Alternatively, there might be an indirect relation between EF and ToM, as posited by Hughes (1998b; Hughes et al., 1998). Children with better EF skills also are likely to have good social and communication skills and thus to have more opportunities for observing social interaction and learning about other people's minds (for recent work along these lines, see Carpendale & Lewis, in press). Our study did not include measures of social interaction quality. Nevertheless, the results suggest that the ability to communicate about internal states is important in the organization of both self-control and the ability to reflect on the mental states of others. Most likely, however, there are multiple pathways to a mature ToM. The notion of "equifinality" has been dominant in the field of developmental psychopathology (e.g., Cicchetti & Rogosch, 1996), but it also may be apt for our understanding of typical development. Do all roads lead to Rome when it comes to ToM? Our research suggests that well-tread paths involving self-control and mental-representational understanding tend to intersect along this journey.

References

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

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin, 121,* 65–94.

Berk, L. E. (1992). Children's private speech: An overview of theory and the status of research. In R. M. Diaz & L. E. Berk (Eds.), *Private speech: From social interaction to self-regulation* (pp. 17–53). Hillsdale, NJ: Erlbaum.

Bretherton, I., & Beeghly, M. (1982). Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology, 18*, 906–921.

Brooks, P. J., Hanauer, J. B., Padowska, B., & Rosman, H. (2003). The role of selective attention in preschoolers' rule use in a novel dimensional card sort. *Cognitive Development, 18*, 195–215.

Carlson, S. M. (2003). Executive function in context: Development, measurement, theory and experience. *Monographs of the Society for Research in Child Development, 68*(Serial No. 274), 138–151.

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

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development, 11*, 73–92.

Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development, 69*, 672–691.

Carpendale, J. I. M., & Lewis, C. (in press). Constructing an understanding of mind: The development of children's social understanding within social interaction. *Behavioral and Brain Sciences*.

Carruthers, P. (1996). Autism as mindblindness: An elaboration and partial defence. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 257–273). Cambridge, England: Cambridge University Press.

Cicchetti, D., & Rogosch, F. A. (1996). Equifinality and multifinality in developmental psychopathology. *Development and Psychopathology, 8,* 597–600.

de Villiers, J., & de Villiers, P. (2000). Linguistic determinism and the understanding of false beliefs. In P. Mitchell & K. J. Riggs (Eds.), *Children's reasoning and the mind* (pp. 191–228). Hove, England: Psychology Press.

Diamond, A. (1988). Abilities and neural mechanisms underlying A not B performance. *Child Development, 59*, 523–527.

Diamond, A., Kirkham, N., & Amso, D. (2002). Conditions under which young children can hold two rules in mind and inhibit a prepotent response. *Developmental Psychology, 38*, 352–362.

Duncan, J. (1986). Disorganization of behavior after frontal-lobe damage. *Cognitive Neuropsychology, 3*, 271–290.

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

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

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., et al. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego, CA: Singular Publishing Group.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology, 17*, 99–103.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1986). Development of knowledge about the appearance–reality distinction. *Monographs of the Society for Research in Child Development*, *51*(Serial No. 212).

Flavell, J. H., Green, F. L., & Flavell, E. R. (1989). Young children's ability to differentiate appearance–reality and Level 2 perspectives in the tactile modality. *Child Development, 60*, 201–213.

Flavell, J. H., & Miller, P. H. (1998). Social cognition. In W. Damon (Series Ed.) & D. Kuhn & R. Siegler (Vol. Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (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.

Frye, D. (2000). Theory of mind, domain specificity, and reasoning. In P. Mitchell & K. J. Riggs (Eds.), *Children's reasoning and the mind* (pp. 149–167). Hove, England: Psychology Press.

Frye, D., Zelazo, P. D., & Burack, J. A. (1998). Cognitive complexity and control: I. Theory of mind in typical and atypical development. *Current Directions in Psychological Science, 7*, 116–121.

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

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of "theory of mind." *Trends in Cognitive Sciences, 7*, 77–83.

Goldsmith, H. H. (1996). Studying temperament via construction of the Toddler Behavior Assessment Questionnaire. *Child Development, 67*, 218–235.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance–reality distinction. *Child Development*, *59*, 26–37.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gopnik, A., Slaughter, V., & Meltzoff, A. N. (1994). Changing your views: How understanding visual perception can lead to a new theory of the mind. In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind: Origins and development* (pp. 157–182). Hove, England: Erlbaum.

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

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 87–114). New York: Wiley.

Griffith, E. M., Pennington, B. F., Wehner, E. A., & Rogers, S. J. (1999). Executive functions in young children with autism. *Child Development, 70*, 817–832.

Hala, S., Hug, S., & Henderson, A. (2003). Executive function and false-belief understanding in preschool children: Two tasks are harder than one. *Journal of Cognition and Development, 4*, 275–298.

Hala, S., & Russell, J. (2001). Executive control within strategic deception: A window on early cognitive development? *Journal of Experimental Child Psychology, 80*, 112–141.

Harris, P. L., & Kavanaugh, R. D. (1993). Young children's understanding of pretense. *Monographs of the Society for Research in Child Development, 58*(Serial No. 231).

Hughes, C. (1998a). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology*, *16*, 233–253.

Hughes, C. (1998b). Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind? *Developmental Psychology*, *34*, 1326–1339.

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

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

Hughes, C., & Russell, J. (1993). Autistic children's difficulty with mental disengagement from an object: Its implications for theories of autism. *Developmental Psychology, 29*, 498–510.

Kloo, D., & Perner, J. (2003). Training transfer between card sorting and false belief understanding: Helping children apply conflicting descriptions. *Child Development, 74*, 1823–1839.

Kochanska, G., Coy, K. C., & Murray, K. T. (2001). The development of self-regulation in the first four years of life. *Child Development, 72*, 1091–1111.

Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology, 36*, 220–232.

Kochanska, G., Murray, K. T., Jacques, T. Y., Koenig, A. L., & Vandegeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development*, *67*, 490–507.

Korkman, M., Kirk, U., & Kemp, S. (1998). *NEPSY: A developmental neuropsychological assessment manual*. San Antonio, TX: Psychological Corporation.

Lewis, C., Freeman, N. H., Kyriakidou, C., Maridaki-Kassotaki, K., & Berridge, D. M. (1996). Social influences on false belief access: Specific sibling influences or general apprenticeship? *Child Development, 67,* 2930–2947.

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

Luria, A. R., Pribram, K. H., & Homskaya, E. D. (1964). An experimental analysis of the behavioral disturbance produced by a left frontal arachnoidal endothelioma (meningioma). *Neuropsychologia, 2*, 257–280.

Meins, E., Fernyhough, C., Wainwright, R., Das Gupta, M., Fradley, E., & Tuckey, M. (2002). Maternal mind-mindedness and attachment security as predictors of theory of mind understanding. *Child Development, 73*, 1715–1726.

Meltzoff, A. N. (1995). Understanding the intentions of others: Reenactment of intended acts by 18-month-old children. *Developmental Psychology, 31*, 838–850.

Meltzoff, A. N., & Decety, J. (2003). What imitation tells us about social cognition: A rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society: Biological Sciences, 358,* 491–500.

Mischel, W., Shoda, Y., & Rodriguez, M. L. (1989). Delay of gratification in children. *Science, 244*, 933–938.

Moore, C., Pure, K., & Furrow, D. (1990). Children's understanding of the modal expression of certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development, 61*, 722–730.

Moses, L. J. (1993). Young children's understanding of belief constraints on intention. *Cognitive Development, 8,* 1–25.

Moses, L. J. (2001). Executive accounts of theory-of-mind development. *Child Development*, *72*, 688–690.

Moses, L. J., & Carlson, S. M. (in press). Self-regulation and children's theories of mind. In M. Chandler, C. Lalonde, & C. Lightfoot (Eds.), *Changing conceptions of psychological life*. Mahwah, NJ: Erlbaum.

Ozonoff, S., Pennington, B. F., & Rogers, S. J. (1991). Executive function deficits in high-functioning autistic individuals: Relationship to theory of mind. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 32*, 1081–1105.

Pears, K. C., & Moses, L. J. (2003). Demographics, parenting, and theory of mind in preschool children. *Social Development, 12*, 1–19.

Perner, J. (1991). *Understanding the representational mind.* Cambridge, MA: MIT Press.

Perner, J. (1998). The meta-intentional nature of executive functions and theory of mind. In P. Carruthers & J. Boucher (Eds.), *Language and thought* (pp. 270–283). Cambridge, England: Cambridge University Press.

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences, 3*, 337–344.

Perner, J., & Lang, B. (2000). Theory of mind and executive function: Is there a developmental relationship? In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from autism and developmental cognitive neuroscience* (pp. 150–181). Oxford, England: Oxford University Press.

Perner, J., & Lang, B. (2002). What causes 3-year-olds' difficulty on the dimensional change card sorting task? *Infant and Child Development, 11*, 93–105.

Perner, J., Lang, B., & Kloo, D. (2002). Theory of mind and self-control: More than a common problem of inhibition. *Child Development*, *73*, 752–767.

Perner, J., Leekam, S. R., & 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.

Putnam, S., & Rothbart, M. K. (2001). *Validation of a Short Form of the Children's Behavior Questionnaire*. Unpublished manuscript, Brandeis University.

Reed, M. A., Pien, D. L., & Rothbart, M. K. (1984). Inhibitory self-control in preschool children. *Merrill-Palmer Quarterly*, *30*, 131–147.

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

Roberts, R. J., Jr., & Pennington, B. F. (1996). An interactive framework for examining prefrontal cognitive processes. *Developmental Neuropsychology*, *12*, 105–126.

Ruffman, T., Perner, J., Naito, M., Parkin, L., & Clements, W. A. (1998). Older (but not younger) siblings facilitate false belief understanding. *Developmental Psychology, 34*, 161–174.

Ruffman, T., Perner, J., & Parkin, L. (1999). How parenting style affects false belief understanding. *Social Development, 8*, 395–411.

Ruffman, T., Slade, L., & Crowe, E. (2002). The relations between children's and mothers' mental state language and theory-of-mind understanding. *Child Development, 73,* 734–751.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94*, 18–38.

Russell, J. (1996). *Agency: Its role in mental development.* Hove, England: Erlbaum.

Russell, J., Jarrold, C., & Potel, D. (1994). What makes strategic deception difficult for children—the deception or the strategy? *British Journal of Developmental Psychology, 12*, 301–314.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87–124.

Siegal, M., & Varley, R. (2002). Neural systems involved in "theory of mind." *Nature Reviews Neuroscience, 3*, 463–471.

Smetana, J. G., Kochanska, G., & Chuang, S. (2000). Mothers' conceptions of everyday rules for young toddlers: A longitudinal investigation. *Merrill-Palmer Quarterly, 46*, 391–416.

Tabachnick, B. G., & Fidell, L. S. (1983). *Using multivariate statistics*. New York: Harper & Row.

Tager-Flusberg, H., & Sullivan, K. (2000). A componential view of theory of mind: Evidence from Williams syndrome. *Cognition, 76*, 59–89.

Taylor, M., & Carlson, S. M. (1997). The relation between individual differences in fantasy and theory of mind. *Child Development, 68*, 436–455.

Vaughn, B. E., Kopp, C. B., & Krakow, J. B. (1984). The emergence and consolidation of self-control from 18 to 30 months of age: Normative trends and individual differences. *Child Development, 55*, 990–1004.

Vygotsky, L. (1962). *Thought and language.* Cambridge, MA: MIT Press.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

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

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development, 72*, 655–684.

Wimmer, H., & Hartl, M. (1991). Against the Cartesian view on mind: Young children's difficulty with own false beliefs. *British Journal of Developmental Psychology, 9*, 125–138.

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.

Zelazo, P. D. (1999). Language, levels of consciousness, and the development of intentional action. In P. D. Zelazo, J. W. Astington, & D. R. Olson (Eds.), *Developing theories of intention: Social understanding and self-control* (pp. 95–117). Mahwah, NJ: Erlbaum.

Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of General Psychology, 1*, 1–29.

Zelazo, P. D., Müller, U., Frye, D., & Marcovitch, S. (2003). The development of executive function in early childhood. *Monographs of the Society for Research in Child Development, 68*(Serial No. 274).

Zelazo, P. D., Reznick, J. S., & Spinazolla, J. (1998). Representational flexibility and response control in a multistep multilocation search task. *Developmental Psychology, 34*, 203–214.

Ziv, M., & Frye, D. (2003). The relation between desire and false belief in children's theory of mind: No satisfaction? *Developmental Psychology, 39*, 859–876.

Received August 14, 2003
Revision received May 26, 2004
Accepted June 23, 2004 ■