Cognitive Development 39 (2016) 71–85

Contents lists available at ScienceDirect

Cognitive Development

A role for executive functions in explanatory understanding of the physical world

Igor Bascandziev[a,*], Lindsey J. Powell[b], Paul L. Harris[a], Susan Carey[a]

[a] Harvard University, United States
[b] Massachusetts Institute of Technology, United States

ARTICLE INFO

Article history:
Received 10 July 2015
Received in revised form 5 February 2016
Accepted 6 April 2016
Available online 14 April 2016

Keywords:
Learning
Naïve physics
Executive functions

ABSTRACT

Are executive functions needed *only* for the *expression* of an already present understanding of the physical world or they are needed for the *construction* of that understanding? We addressed this question in the context of Hood's (1995) tubes task. When asked to find a ball dropped down an opaque curved tube, 2- and 3-year-olds search *directly* below the place where they have seen the ball dropped, rather than at the bottom of the *tube* into which the ball was dropped. Instructions about the role of the tubes, but not visual feedback about the correct location of the ball, help children improve on this task. We found that children who scored higher on specific EF measures and on performance IQ showed greater improvement on the tubes task after receiving instructions about the role of the tubes than did children with lower EFs and performance IQ. These results suggest that EFs are needed for the *construction* of new explanatory understanding of the physical world.

1. Introduction

Our conceptual understanding of the world undergoes profound changes in early childhood, due to both the addition of new factual knowledge to our existing theories about the world and, more rarely, overall structural changes in those theories. The timing and nature of concept addition and change vary substantially from one domain to the next. Therefore, conceptual development must first be examined within the domain of a particular theory. However, a full understanding of conceptual development also requires investigating the role of *general* cognitive capacities in learning *across* different domains. Here we investigate the contribution of executive functions (EFs) to preschool children's ability to change their understanding of the physical world.

EFs are a suite of abilities that include holding and flexibly manipulating thoughts in mind, maintaining a particular goal while inhibiting various types of endogenous and exogenous interference, and selecting appropriate responses. Three core EFs have been identified in the literature: (a) inhibition, (b) working memory, and (c) set shifting (Diamond, 2013; Miyake et al., 2000). These core EFs are in turn thought to underlie more complex cognitive reasoning skills, such as comprehension and consistency monitoring, the formation of abstract representations, and the construction and maintenance of the hierarchical rules that guide behavior (Diamond, 2013; Kharitonova & Munakata, 2011; Snyder & Munakata, 2010; Zelazo & Frye, 1997), and they are part of the construct of fluid IQ (Duncan, Burgess, & Emslie, 1995).

* Corresponding author.
E-mail address: igb078@mail.harvard.edu (I. Bascandziev).

http://dx.doi.org/10.1016/j.cogdev.2016.04.001
0885-2014/

EFs are certainly linked to learning in childhood, as they correlate with and predict academic achievement. EFs are more strongly associated with school readiness than is IQ, entry-level reading skills, or entry-level math skills (Blair & Razza, 2007; Diamond, Barnett, Thomas, & Munro, 2007; Morrison, Ponitz, & McClelland, 2010). Moreover, EFs maintain their importance throughout the school years; indeed, working memory and inhibition independently predict math and reading scores in every grade from preschool through high school (e.g., Blair & Razza, 2007; Gathercole, Tiffany, Briscoe, Thorn, & ALSPAC Team, 2005). There are many reasons this relationship might hold, including, possibly, that EFs are crucial for the acquisition of conceptual knowledge. The present paper begins to explore this possibility in the context of developing an understanding of the physical world, building on parallels with the much more fully explored case study of the relations between EFs and developing theory of mind (ToM).

1.1. Case study of relations between EFs and developing ToM

A large literature on early developing ToM supports five generalizations. First, young infants have rich abstract representations of intentional agents. Infants distinguish self-moving objects from dispositionally inert ones, and reason about actions of self-moving objects in terms of those objects' goals and perceptual states (Gergely & Csibra, 2003; Luo & Baillargeon, 2007; Luo & Johnson, 2009; Woodward, 1998). By the second year of life at the latest, infants predict these agents' future actions in terms of the information those other agents have had access to, as if tracking their epistemic states (Kovács, Téglás, & Endress, 2010; Onishi & Baillargeon, 2005; Southgate, Senju, & Csibra, 2007; Surian, Caldi, & Sperber, 2007; see Carey, 2009 for review). A second generalization is that despite such representations being attested in infancy, when preschool children are asked to explicitly access these representations they robustly fail to manifest knowledge with the same content. The striking failures of children under age 4 on explicit false belief tasks are well known (Wellman, Cross, & Watson, 2001), and are interrelated with failures at articulating sources of knowledge (Gopnik & Graff, 1988; O'Neill & Gopnik, 1991), the appearance/reality distinction (Flavell, Green, & Flavell, 1986; Gopnik & Astington, 1988), and Level II perspective taking (Flavell, Everett, Croft, & Flavell, 1981; Flavell, Green, & Flavell, 1989; Moll & Meltzoff, 2011).

There are two broad, possible accounts of preschoolers' failures in the face of infants' success. One posits *continuity* from infancy throughout development in representational content, and explains preschoolers' failure in terms of performance demands in the preschool tests that are lacking in the infant tests. In particular, it has been suggested that the preschool tasks make executive function (EF) demands that the infant tasks do not, frequently because the preschool tasks are designed to pit correct answers against incorrect ones that align with good heuristic guides (e.g., people usually look for objects where they are because they usually have true beliefs about the objects' locations). Suppressing answers with this intuitive appeal plausibly draws upon children's inhibitory control. On this account, the developmental changes observed in the preschool years reflect known developmental advances in executive function, rather than changes in the representations that underlie reasoning about agents and their mental states. The third generalization from the ToM literature is that, indeed, measures of EF predict performance on explicit ToM tasks in the preschool years. A recent meta-analysis of data from one hundred studies conducted over the last 20 years in 15 different countries included almost 10,000 3–6-year-old participants and confirmed a significant relationship between performance on preschool false belief tasks and EFs, especially measures of conflict inhibition, even after controlling for age and verbal ability (Devine & Hughes, 2014). Finally, as continuity theory would predict, temporary depletion of EFs leads to decreased performance on false-belief tasks in 4- and 5-year-olds (Powell & Carey, 2016).

While not denying that EFs are likely critical to performance on tests of ToM, and that EF development may thus account for some of the improvements on tests of ToM understanding with age (Moses, 2001), the *fourth* generalization from the ToM literature is that there is *also* learning-driven change within the domain-specific representations and computations children bring to the tasks. That is, ToM development requires *learning/construction* of new representational resources during the preschool years. Performance on a variety of preschool ToM tasks is influenced by exposure to input illustrating the link between thoughts and behaviors, including input from explicit training (Appleton & Reddy, 1996; Slaughter & Gopnik, 1996), amount of mental state language in parental input (Ruffman, Slade, & Crowe, 2002), and environmental factors such as having older siblings (Perner, Ruffman, & Leekam, 1994). Finally, Sabbagh, Xu, Carlson, Moses, and Lee (2006) compared over 100Chinese with over 100 American 3.5–4.5-year-olds on large batteries of EF measures and ToM measures. In this sample, the Chinese children were a full 6 months ahead of the American children on all of the EF measures, but the two populations were identical in their performance on the ToM tasks. Thus, superior EFs are not sufficient for superior performance on ToM tasks. This result undermines the claim that the maturation of EFs can completely explain the developmental changes on performance on ToM tasks, contrary to the continuity hypothesis. Some learning or construction specific to ToM is also implicated in the observed developmental changes in the preschool years.

EFs may be implicated, as suggested above, in the *expression* of conceptual understanding in the context of a particular task. Researchers who accept the continuity assumption hold that the development of the capacity to *express* the innate knowledge is the *full* explanation of the developmental changes observed in the preschool years on the explicit tasks, and also the *full* explanation of the correlations between measures of young children's EF and their performance on the explicit tasks. We call this the *expression alone* hypothesis. Contrary to the *expression alone* hypothesis, the *fifth* generalization from the ToM literature is that EFs are drawn upon by the learning/construction mechanisms that underlie change in ToM in the preschool years. One source of evidence for this generalization derives from longitudinal data, which show that variation in early measurements of EFs predict later performance on ToM tasks rather (or more) than the reverse (Carlson, Mandell, &

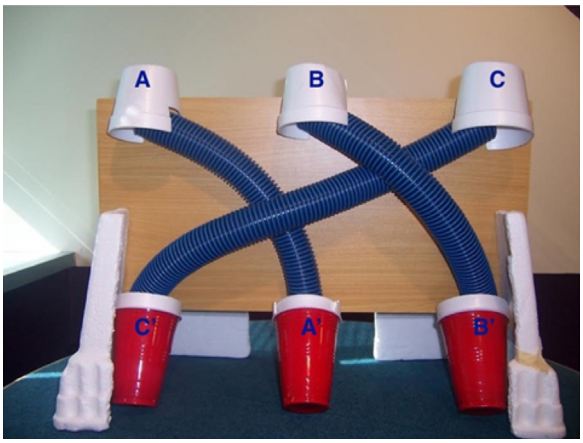

Fig. 1. Picture of the tubes apparatus.

Williams, 2004; Hughes & Ensor, 2007; Hughes, 1998). Such a pattern suggests that EFs play a role in theory change across the intervening period, rather than solely contributing to performance during the later assessment. A related research strategy is to provide children with training on a particular task, eliciting learning over the course of the experiment, and to test the relationship of this learning to children's EFs. Insofar as conceptual learning or change demands EFs, children with stronger EFs should be better able to take advantage of exposure to relevant information in the training than children with weaker EFs. One study has adopted this strategy, finding that 3.5-year-old children who failed a false belief pretest were better able to benefit from false belief training the stronger their EF skills were (Benson, Sabbagh, Carlson, & Zelazo, 2013).

1.2. Parallels with a case study in the domain of intuitive physics

While much less studied, data on infants' and preschoolers' concepts of the physical world support four of the five generalizations about the ToM case study detailed above. *First,* infants represent bounded, coherent, separately movable objects as spatiotemporally continuous, characterized by solidity such that one cannot move through the space occupied by another, and subject to the constraints of Michottian contact causality (see Baillargeon, Spelke, & Wasserman, 1985; Leslie & Keeble, 1987; Leslie, 1982; Spelke, Keestenbaum, Simons, & Wein, 1995; see Carey, 2009 for review). *Second,* in the face of evidence for rich abstract knowledge of objects and their physical interactions, two tasks (the tubes task and the doors task) reveal robust failures of preschool children to demonstrate knowledge of the physical world that is attested in young infants.

In the tubes task, young children commit a gravity error when asked to find a ball dropped down one of the curved opaque tubes connecting three chimneys above with three containers below (see Fig. 1). Instead of following the shape of the tube in which the ball was dropped, 2- and 3-year-olds persistently search in the container that is directly below the chimney in which they have seen the ball disappear, as if they expect gravity to make objects fall straight down regardless of other constraints (e.g., when a ball is dropped in A, they search in C' rather than A'; Hood, 1995). Thus, young preschool aged children do not seem to use their knowledge about solidity and continuity to guide their search in this task. By age 4–4.5 years, children no longer commit the gravity error when they are presented with a single ball dropped into one of the tubes in an apparatus such as that of Fig. 1 (Hood, 1995). (See Berthier, DeBlois, Poirier, Novak, & Clifton (2000) and Hood, Carey and Prasada (2000) for convergent evidence for this conclusion from the doors task, which is more closely modeled on the infant paradigms in which infants as young as 2 months of age display such knowledge).

Third, consistent with the *continuity* of representations hypothesis, lowering EF demands leads to improved performance and increasing EF demands leads to worse performance on the tubes task. Lee and Kuhlmeirer (2013) compared eye-gaze behavior – which does not require explicit answers and therefore does not have the same EF demands as a search task – with the pointing behavior of 26-month-olds in the tubes task. There were two levels of complexity in this study: (a) one tube connecting one of three chimneys above with one of three containers below, while the other chimneys and containers remained unconnected, and (b) two tubes, each connecting one of the three chimneys above with one of the three containers below. Children gazed into the correct container but then pointed to the incorrect gravity container in the one-tube configuration. However, most children did not exhibit such behavior in the two-tube configuration. That is, most children both gazed and pointed toward the gravity container. In conclusion, the correct eye-gaze behavior in the one-tube configuration suggests that under some circumstances, young children do have an implicit knowledge about the true location of the ball.

Conversely, Hood, Wilson and Dyson (2006) presented 4.5-year-olds who were no longer committing the gravity error with the task of following two balls instead of one on the tubes task. As predicted, these children reverted to committing the gravity error. A plausible interpretation of this finding is that although these children had the competence to perform well

on this task, taxing their EFs (working memory, inhibition of two gravity responses) resulted in masking that competence, which led to the reappearance of the gravity error.

Finally, Baker, Gjersoe, Sibielska-Woch, Leslie, and Hood (2011) showed that EFs are correlated with performance on the tubes and the doors tasks. Specifically, children with higher delay inhibitory control performed better on both the doors task and the tubes task, most likely because children who failed to inhibit their urge to start searching as soon as possible, searched incorrectly.

That the tubes task taxes EFs in young children does not preclude the possibility that domain-specific conceptual development *also* has a role to play in the disappearance of the gravity error. Many findings show that the *fourth* generalization from the ToM literature applies to this case study as well. Learning/construction, as well as EF development, plays a role in driving the dramatic developmental changes between ages 2 and 4. Children must construct an explicit model of the causal role of the tubes that allows them to correctly compute the correct location of the ball regardless of the shape, configuration, or complexity of the intertwined tubes. That children do not initially have such model was shown in a study in which 2-year-olds were presented with an apparatus that was tilted by 90° so that the chimneys, the containers, and the tubes were horizontal. This change removed the inhibitory control demands from this task, because the ball was not falling and thus could not have elicited a prepotent hypothesis that *falling* objects fall in a straight line. This change eliminated young children's tendency to favor the cup directly in line with the starting location when searching for the ball. Yet children performed no better than chance (Hood, Santos, & Fieselman, 2000). In addition to being biased to expect that objects will fall in a straight line, young children also do not understand how the tubes constrain the movement of the ball.

A salient finding from the literature on the tubes task is that even though children search repeatedly, sometimes as many as 10 trials, with feedback on the correct location of the ball, they never stop making the gravity error (Bascandziev & Harris, 2010; Hood, 1995). In contrast, other studies show that children improve dramatically when they are provided with information about the causal role of the tubes. For example, in two recent intervention studies, children were first given pretest trials, and they received verbal instruction about the mechanism through which the tubes constrain the motion of the ball. Finally, they were given posttest trials. In contrast to control instructions, instructions targeting children's understanding of the tubes mechanism significantly improved their performance on posttest (Bascandziev & Harris, 2010; see also Bascandziev & Harris, 2011, for convergent evidence that making the causal role of the tubes more salient improves performance). In a different study, children were asked on each trial to imagine the ball inside the tube before the ball was dropped into it. This request led to fewer gravity errors (Joh, Jaswal, & Keen, 2011).

Thus, parallel to the ToM domain, in the domain of knowledge of objects and their physical interactions, young preschoolers fail to display knowledge young infants display in looking time studies, their failures are sensitive to manipulations of EF demands of the tasks and their likelihood of failure are correlated with measures of EF. And contrary to the *expression alone* hypothesis concerning the source of developmental change on the tubes task during the preschool years, there is evidence that learning/construction also plays a role. Accepting this conclusion raises the possibility that EFs may also play a role in that learning or construction process. Notice that accepting the learning/construction hypothesis (as opposed to the continuity hypothesis/expression alone hypothesis), does not by itself, favor the hypothesis that EFs are drawn upon during learning. It depends what the nature of the learning process is. Many possible learning mechanisms (e.g., associative learning) plausibly do not draw heavily upon EF resources. However, many plausibly do, especially those that implicate comprehension monitoring and explicit conflict detection, which may play an important role in recognizing when one's theory falls short of explaining the world and is in need of extension or alteration. The goal of the present study is to test the *fifth* generalization that has emerged from the ToM literature in the domain of physical reasoning. Namely we test the hypothesis that EFs are drawn upon in the learning process that contributes to developmental changes on the tubes task.

1.3. The present study

Our study takes the approach of measuring individual differences in EF and assessing their relationship to learning over the course of training. We assessed children's pretest performance on the tubes task, followed by training that involved verbal instruction, followed by a posttest on the tubes task. In addition, children were tested on an EF battery consisting of (1) two response conflict inhibitory control/set shifting tasks that require from children to inhibit a prepotent response, to select a response that is in conflict with the prepotent response, and to shift between sets of rules, (2) two delay inhibitory control tasks that require from children only to inhibit a prepotent response, and (3) two working memory tasks. We also administered a performance IQ test, to measure children's non-verbal fluid intelligence. Previous research has found that tests developed to measure fluid intelligence are highly correlated with tests designed to measure core EFs (Burgess, Gray, Conway, & Braver, 2011; Conway, Kane, & Engle, 2003; Duncan et al., 1995; Engel de Abreu, Conway, & Gathercole, 2010; Friedman et al., 2006; Fry & Hale, 2000; Fukuda, Vogel, Mayr, & Awh, 2011). Finally, we administered a receptive vocabulary task (a measure of crystallized intelligence) in order to control for children's verbal abilities. If the construction of a model of the causal role of the tubes draws on EFs and/or PIQ, then children who score higher on EFs/PIQ should show steeper learning slopes, indicating greater improvement on the tubes task from pretest to posttest, relative to children who score lower on EFs/PIQ, even when age and receptive vocabulary are controlled for.

2. Method

2.1. Participants

The sample included 100 children (M_{age} = 38 months, SD = 2.34, range = 34–42 months, 56 females). An additional five children were tested, (M_{age} = 37 months), but were excluded because they did not follow the directions or did not complete the tasks.

2.2. Constructs and tasks

2.2.1. The tubes task

Pre- and posttest scores on this task comprised the dependent variables in this study. Of particular interest was the *improvement* between pretest and posttest performance. In this task, the experimenter familiarized children with the apparatus (see Fig. 1) and the task. Each child then received seven pretest trials. The ball was dropped into different tubes according to the following predetermined sequence: C, A, B, A, B, C, A. Children were invited to search for the ball after each drop. Children received visible feedback about the correct location of the ball on each trial, but no verbal feedback. If they did not continue searching until they found the ball, the experimenter revealed its correct location. Following the pretest and regardless of their performance, children received verbal instruction during two training trials. After dropping the ball in the middle tube (tube B) on each training trial, the experimenter said, "Look! I dropped the ball in this tube. This is the top of the tube. Can you show me where the bottom of this tube is? That's right. This is the top and this is the bottom of the tube. And you know what? The ball could not escape from this tube. It rolled inside this tube into this cup." At that point, the experimenter retrieved the ball from the cup. After the two training trials, children received seven posttest trials in which the ball was dropped into tubes according to the following sequence: C, A, C, A, B, C, A. Children again received visible feedback throughout, but were not given any instructions or verbal feedback during the posttest trials. Note that the training tube B was used only on one trial at posttest.

2.2.2. Response conflict inhibitory control

The Dimensional Change Card Sort and Day-Night task assessed this construct. These tasks are traditionally used to measure conflict inhibitory control (both tasks) and set shifting (DCCS task). In the Dimensional Change Card Sort task, children were presented with pictures that varied along two dimensions (e.g., shape and color). Children first sorted a set of cards along one dimension, and were then asked to switch and sort them into the same containers but along the other dimension (see Frye, Zelazo, & Palfai, 1995 for details). There were five post-switch trials and so the possible range of scores was between 0 and 5. In the Day-Night task, children saw pictures of a day sky and a night sky and were asked to engage in a series of incongruent pointing trials, where the experimenter provides a label ("Day" or "Night") and the child points to the picture that does not match that label (see Gerstadt, Hong, & Diamond, 1994 for details). Children completed 10 test trials. Thus, the scores could range between 0 and 10.

2.2.3. Delay inhibitory control

The Gift Delay and Tower tasks assessed this construct. In Gift Delay, the experimenter announced that she had a present for the child. She told the child that the present was a surprise and that the child was not allowed to peek while she wrapped it up. Then the experimenter spent 60 s wrapping the present very noisily behind the child's back. Children received three different scores on this task: (a) how much they turned (4 possible categories adopted from Kochanska, Murray, Jacques, Koenig, & Vandegeest, 1996; 1: turned around and refused to turn back; 2: turned fully, but then turned back to face the camera; 3: turned enough to see the present but not fully; 4: did not turn at all), (b) how long they waited before turning (measured in seconds), and (c) how many times they turned. In the Tower task, the experimenter established that she and the child should take turns placing blocks to make a tower, and then gave the child repeated opportunities to either wait for her turn or continue placing blocks (see Kochanska et al., 1996 for details). Each child built two towers. The scores could range between 0 and 12 and they represented the average number of times children did not wait for their turn while building a tower across the two building sessions.

2.2.4. Working memory

Forward Digit Span and the Self-Ordered Pointing Task assessed children's working memory. In Forward Digit Span, children repeated sequences of numbers in the same order told to them by an experimenter. After each successful trial, the experimenter increased the length of the next sequence by one, until the child failed to accurately repeat two sequences of the same length in a row. In the Self-Ordered-Pointing task, in each trial, the child saw a series of pictures featuring the same set of objects scrambled into different locations. When presented with each picture, the child had to point to an object they had not pointed to in a previous picture[1] (see Hongwanishkul, Happaney, Lee, & Zelazo, 2005 for details). The task was

[1] The stimulus images used in this task courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/.

discontinued when children failed two consecutive times on sets with the same number of pictures. The score each child received equaled the number of pictures presented on the last set that they had successfully passed.

2.2.5. Performance IQ

Two subtests of the Wechsler Primary and Preschool Scale of Intelligence (WPPSI-III) assessed this construct: (a) Block Design, which asks children to arrange white and red blocks so that they match a design presented by the experimenter and (b) Object Assembly, which asks children to reassemble objects cut into increasing numbers of pieces.

2.2.6. Receptive vocabulary

As a measure of crystalized verbal IQ, we administered a subtest of WPPSI-III's verbal IQ test in which children are asked to select one of four pictures that corresponds to a word.

2.3. Procedure

The tasks were presented in two 25-min sessions that were approximately one week apart. The order in which the tasks were presented to children was fixed across participants. Session I included Block Design, Dimensional Change Card Sort, Gift Wrap, and Digit Span, in that order. Session I ended with the pretest trials on the tubes task, followed by training and then posttest trials. Session II consisted of Receptive Vocabulary, Day-Night, Tower, Self-Ordered Pointing task, and Object Assembly, in that order. The tasks were coded offline from the video, and a second, independent coder coded 20% of the final sample. We assessed the inter-rater agreement by calculating intraclass correlations between the coders' scores, which were high, ranging between 0.90 and 1.

3. Results

3.1. Descriptive statistics and preliminary analyses: tubes task

3.1.1. Pretest performance on the tubes task

One hundred children completed the pretest trials. The proportion of correct searches on pretest was $M = 0.44$, $SD = 0.31$. The proportion of incorrect gravity searches was $M = 0.49$, $SD = 0.31$ and the proportion of incorrect searches in the third (i.e., non-gravity) location was $M = 0.07$, $SD = 0.12$. Thus, most errors that children made on pretest were gravity errors. Furthermore, preliminary analysis showed that pretest performance on the tubes task was significantly correlated with age, $r = 0.23$, $p = 0.019$ and gender, $r = 0.38$, $p < 0.0001$. On average, older children and males tended to score better. These findings are consistent with those of previous studies (Bascandziev & Harris, 2011; Hood, 1995).

3.1.2. Improvement after receiving training on the tubes

Ninety-five of the 100 participants completed the posttest trials. Of these 95 children, two did not complete all seven trials on pretest and posttest. These two children were included in the analysis and we therefore computed the proportion of trials on which children searched in the correct cup, incorrect (gravity) cup, or incorrect (non-gravity) cup. Furthermore, given that most scores for the proportion of correct searches fell outside the range 0.3–0.7, we transformed that variable with an arcsine transformation. Note that while the analyses were conducted with the transformed variable, the proportions presented in text are de-transformed scores. The average proportion of correct searches for the 95 children was $M = 0.43$, $SD = 0.31$ on pretest and $M = 0.48$, $SD = 0.37$ on posttest, reflecting a statistically significant improvement, $t(94) = 2.55$, $p = 0.012$, $d = 0.26$. There was also a significant reduction in the average proportion of gravity searches, from $M = 0.49$, $SD = 0.31$ on pretest to $M = 0.44$, $SD = 0.35$ on posttest, $t(94) = -2.27$, $p = 0.026$, $d = 0.23$. The average proportion of non-gravity searches did not change significantly from pretest $M = 0.07$, $SD = 0.12$, to posttest $M = 0.08$, $SD = 0.14$. Thus, the increase in correct searches from pretest to posttest resulted from reduction of gravity errors alone. Importantly, not all children improved from pretest to posttest—some did and some did not. This variability allowed us to explore the ways in which children who improved on the tubes task differed from children who did not.

To assess whether children's performance improved only on the tube on which they had received training (i.e., Tube B, the middle tube), we computed the pretest and posttest average performance for each tube separately. Children were not more likely to show improvement from pretest to posttest on the training tube (pretest $M = 0.57$; posttest $M = 0.59$, $p > 0.1$) than on the other two tubes (Tube A: pretest $M = 0.49$; posttest $M = 0.54$, $p > 0.1$; Tube C: pretest $M = 0.25$; posttest $M = 0.40$, $t(94) = 3.69$, $p < 0.001$, $d = 0.38$). Thus, children showed considerable improvement on the longest and most intertwined tube (tube C), which they also found to be the most difficult to solve at pretest. More importantly, these results suggest that the training did not teach children a simple rule about one particular instance (the middle tube), but it gave them generic knowledge about how the tubes mechanism works.

3.2. Descriptive statistics and composite scores (predictor and control variables)

Panel 1 of Table 1 presents the descriptive statistics of all EF tasks. The scales for Tower and Number of Peeks during Gift Wrap were reversed so that all EF measures were in the same direction (i.e., high scores reflect high EF ability). All EF scores

Table 1
Descriptive statistics for the EF (Panel 1) and the WPPSI (Panel 2) measures.

Panel 1

	Card Sort	Day-Night[a]	Digit Span	Self-Ordered Pointing	Peek Score – Gift Wrap	Number of Peeks – Gift Wrap	Latency – Gift Wrap	Tower
Mean	2.94 (n = 94)	0.47 (n = 97)	3.32 (n = 73)	4.08 (n = 86)	3.53 (n = 99)	0.87 (n = 99)	42.42 (n = 99)	2.94 (n = 89)
Standard Deviation	1.28	0.34	0.86	1.28	0.9	1.38	24.75	3.22
Range	0–5	0–1	0–6	2–6	1–4	0–5	1–60	0–12

Panel 2

	Receptive Vocabulary Raw	Receptive Vocabulary Scaled	Block Design Raw	Block Design Scaled	Object Assembly Raw	Object Assembly Scaled	Performance IQ Scaled
Mean	20.41 (n = 93)	13.3[b] (n = 93)	17.74 (n = 94)	11.61[b] (n = 94)	11.88 (n = 89)	11.79[b] (n = 89)	109.43[b] (n = 84)
Standard Deviation	5.04	2.55	3.03	2.26	6.06	2.71	11.56
Range	10–31	8–19	4–26	4–18	2–28	2–17	79–134

[a] Because some children did not complete all ten trials on the day-night task, we computed the proportion of trials on which the child pointed at the correct picture.
[b] The scaled Receptive Vocabulary mean was 1.1 standard deviations above the national norm group. The scaled Block Design mean was 0.54 standard deviations above the national norm group. The scaled Object Assembly mean was 0.6 standard deviations above the national norm mean. The composite performance IQ mean was 0.63 standard deviations above the national norm mean.

were standardized with a mean of 0 and a standard deviation of 1. This allowed us to aggregate the three Gift Wrap measures by computing an average across the three scores (Peek, Number of Peeks, and Latency), which were highly inter-correlated (*r*s ranging from 0.76 to 0.90, Cronbach's Alpha = 0.94). This produced a single composite Gift Wrap variable.

Panel 2 of Table 1 presents the descriptive statistics of the raw and the scaled scores on WPSSI–III, the latter adjusted to account for the age of the participant. In contrast, all EF scores in our study are raw scores that have not been adjusted to account for the age of the participant, and some of the variance in the EF scores is due to age differences. Therefore, in order to have comparable predictor and control variables, in our analyses we used the raw WPPSI-III scores. As with the EF measures, these scores were standardized to have a mean of 0 and a standard deviation of 1.

Previous research has suggested that response conflict inhibitory tasks (RC-EF henceforth) and delay inhibitory tasks (D-EF henceforth) tap into different underlying constructs (Carlson & Moses, 2001; Hongwanishkul et al., 2005). Whereas RC-EF tasks require children to both inhibit a prepotent response and to select a response that is in conflict with the prepotent response, D-EF tasks require children only to inhibit a prepotent response. Our correlational analyses confirm that different constructs are involved. There were small to moderate partial correlations (controlling for age, gender, and receptive vocabulary) between the two RC-EF measures (Card Sorting and Day Night; $r = 0.19$, $p = 0.09$) and between the two D-EF measures (Gift Wrap and Tower; $r = 0.29$, $p = 0.007$). At the same time, the partial correlations between the RC-EF measures and the D-EF measures were weak and not significant (*r*s ranging between 0.01–0.16, *p*s > 0.1). For this reason, we constructed an RC-EF inhibitory composite variable by aggregating the two conflict inhibitory tasks (Cronbach's Alpha = 0.31) and a D-EF composite variable by aggregating the two delay inhibitory tasks (Cronbach's Alpha = 0.48). The composite variables were then standardized with a mean of 0 and a standard deviation of 1.

The partial correlation (controlling for age, gender, and receptive vocabulary) between the two working memory measures, digit span (DSPAN henceforth) and self-ordered pointing task (SOP henceforth) was weak and not significant ($r = 0.07$, $p = 0.6$), suggesting that these tasks tap into different working memory structures. Specifically, the digit span task requires children to remember verbal information, whereas the self-ordered pointing task requires children to remember the appearance of the objects presented in pictures. The lack of a significant correlation between these two tasks is consistent with the proposal that working memory is composed of several independent components among which are the phonological loop, a buffer responsible for holding verbal information, and the visuospatial sketchpad, a buffer responsible for holding visuospatial information (Baddeley & Hitch, 1974; Baddeley, 2000; Gathercole, Pickering, Ambridge, & Wearing, 2004). This theoretical consideration, plus the fact that the working memory tasks were not correlated with each other, led us to keep SOP and DSPAN as separate measures.

As anticipated, the partial correlation (controlling for age and gender) between the two performance IQ subtests (object assembly and block design) was statistically significant ($r = 0.30$, $p = 0.006$). This result suggests that the two performance IQ (PIQ) subtests tap into the same construct, and so we constructed a PIQ aggregate score. This variable was then standardized with a mean of 0 and a standard deviation of 1. The partial correlations (controlling for age and gender) between block design (PIQ subtest) and receptive vocabulary, between object assembly (PIQ subtest) and receptive vocabulary, and between the composite PIQ and receptive vocabulary were not statistically significant (*r*s ranging between 0.08–0.15, *p*s > 0.1), consistent

Table 2
Bivariate (above diagonal) and partial correlations (below diagonal, controlling for Age, Gender, and Receptive Vocabulary [raw scores]) among predictor variables and pretest performance on the tubes task.

	PreTest Tubes	RC-EF	D-EF	PIQ	DSPAN	SOP
PreTestTubes	*	0.10 (0.32)	0.10 (0.37)	**0.37 (<0.001)**	0.09 (0.45)	**0.23 (0.03)**
RC-EF	0.14 (0.21)	*	0.06 (0.57)	**0.33 (0.004)**	**0.25 (0.04)**	**0.22 (0.06)**
D-EF	**0.19 (0.08)**	0.04 (0.73)	*	**0.25 (0.03)**	**0.38 (0.002)**	0.007 (0.95)
PIQ	**0.29 (0.01)**	**0.29 (0.01)**	**0.31 (0.007)**	*	**0.26 (0.04)**	**0.23 (0.04)**
DSPAN	**0.21 (0.09)**	**0.28 (0.03)**	**0.35 (0.006)**	**0.29 (0.02)**	*	0.09 (0.50)
SOP	0.16 (0.14)	0.19 (0.10)	0.03 (0.78)	0.12 (0.30)	0.07 (0.60)	*

Notes: p values shown in parentheses. Partial correlations below diagonal. Pretest Tubes Arcsine transformed. Significant and trends toward significant correlations in bold. As noted in Table 1 above, the number of children who completed each task ranged between 73 and 99. Preliminary analyses were consistent with the assumption that these data were missing completely at random (MCAR). Thus, we used listwise deletion and so the number of participants in the analyses presented here ranged between 61 and 91.

with the conclusion that receptive vocabulary taps into a different construct from PIQ and should therefore be kept as a separate variable.

3.2.1. Relationship between EFs and WPPSI-III

The correlations between receptive vocabulary and EF measures (RC-EF, D-EF, and each working memory measure) were weak and not statistically significant (rs ranging between −0.007 to 0.18, $ps > 0.1$). This is in line with the proposal that receptive vocabulary is a measure of crystallized intelligence, which is separate from fluid intelligence and independent from EFs. Conversely, PIQ was correlated with all EF measures (see Table 2). Moreover, these correlations (with the exception of the correlation between PIQ and SOP) remained significant even after partialling out age, gender, and receptive vocabulary. This result is consistent with previous research showing a strong relationship between fluid intelligence and core EFs (e.g. Burgess et al., 2011; Conway et al., 2003; Duncan et al., 1995; Engel de Abreu et al., 2010; Friedman et al., 2006; Fry & Hale, 2000; Fukuda et al., 2011). Moreover, it is consistent with the theoretical proposals mentioned above according to which higher-order EFs, at least partially measured by performance IQ tests, are built out of core EFs.

3.3. Predictor variables (Performance IQ and EFs) and performance on the tubes pretest

The partial correlations reveal a trend toward a significant correlation between pretest performance on the tubes task and D-EF. This finding replicates the finding of Baker et al. (2011) that delay (but not response conflict) inhibitory control measures are correlated with performance on the tubes task. In other words, children who are better at inhibiting impulsive behaviors score higher on the tubes task. Informal observations confirm that children indeed feel the urge to find the ball as fast as possible without stopping to think about its trajectory. Thus, measures that differentiate children who inhibit impulsive behaviors from those that do not seem to be the best predictors of young preschoolers' initial performance on the tubes task.

Further inspection of the partial correlations presented in Table 2 reveals a trend toward a significant correlation between pretest performance on the tubes task and DSPAN and a statistically significant correlation between performance on the tubes task and PIQ. In summary, the results of the present study are consistent with the findings of previous studies. They go beyond those findings in showing that performance on the tubes task is correlated with fluid intelligence and with working memory.

3.4. Predictor variables (Performance IQ and EFs) and the rate of change on the tubes task

Our main research question is whether individual differences in EF and PIQ are associated with the rate of change on the tubes task. The nature of our data (repeated measures on the dependent variable and continuous predictor and control variables) allowed us to test our hypothesis by fitting multilevel models for change to data. There are two levels in the model. The level 1 submodel describes the pretest scores and the rate of change on the tubes task. The level 2 submodel, specifies the relationship between predictor variables and children's pretest scores and rates of change on the tubes task. Appendix A provides a mathematical description of this multilevel model, and it also presents the results of the various fitted multilevel models. The parameter estimates that describe the relationship between children's pretest performance and predictor variables (see upper panel of Table A1) duplicate the correlational analyses presented in Table 2, and so they are not discussed again. Here, we present the parameter estimates that describe the relationship between predictor variables and the rate of change (see lower panel of Table A1), as this is the focus of our investigation.

3.4.1. Predicting the rate of change

Given that we have created several different predictor variables for EFs, we fitted separate models to the data (one model for each predictor variable). Each of these models controls for age, gender, and receptive vocabulary. The parameter estimate γ_{11} is of primary interest in this study because it tests the hypothesis that EFs and PIQ are associated with the rate of change on the tubes task (see Appendix A). A preliminary analysis showed that there was no significant relationship between D-EF

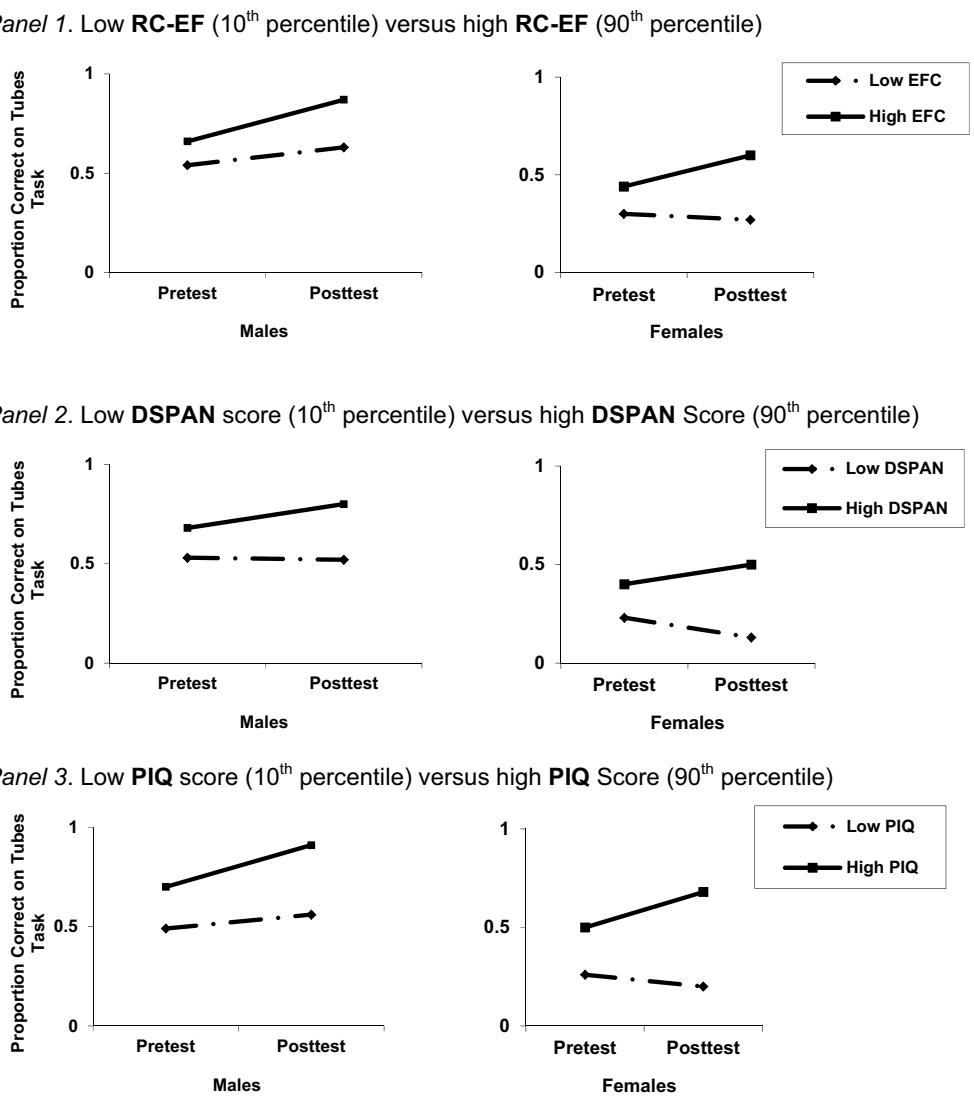

Fig. 2. Fitted trajectories for a prototypical boy and a prototypical girl with an average age, average receptive vocabulary score, a low EF/PIQ score (10th percentile) and a high EF/PIQ score (90th percentile).
Note: The dependent variable is de-transformed and it represents a mean proportion of correct searches on the tubes task.

or SOP and the rate of change on the tubes task ($\gamma_{11} = 0.004$, $p = 0.95$ and $\gamma_{11} = 0.02$, $p = 0.73$ respectively). Therefore, we dropped these two predictors from this analysis.

Controlling for age, gender, and receptive vocabulary, the parameter estimate γ_{11} is statistically significant for DSPAN (tubes rate of change gain of 0.09 for every 1-unit difference in DSPAN) and for PIQ (rate of change gain of 0.12 for every 1-unit difference in PIQ). There is a trend such that the predicted average rate of change in performance on the tubes task is 0.08 higher for every 1-unit difference in RC-EF. None of the other variables (Age, Gender, and Receptive Vocabulary) presented in Table A1 in the Appendix was consistently correlated with children's ability to improve on the tubes task.

Fig. 2 provides a visual representation of the effect of RC-EF, DSPAN, and PIQ on children's improvement on the tubes task. Note that gender is a control variable in the model and given that there is no such thing as a child with an average gender, we present a prototypical boy and a prototypical girl with an average age, average receptive vocabulary and in Panel 1 with a low RC-EF score (10th percentile) versus a high RC-EF score (90th percentile). Panel 2 presents fitted trajectories of a prototypical boy and a prototypical girl with a low DSPAN score (10th percentile) versus a high DSPAN score (90th percentile). Finally, Panel 3 presents fitted trajectories of a prototypical boy and a prototypical girl with a low PIQ score (10th percentile) versus a high PIQ (90th percentile). All three graphs show that, irrespective of gender, children with higher RC-EF, D-SPAN and PIQ improve at a higher rate after receiving testimony about how the tubes constrain the movement of the ball compared to children with lower RC-EF, D-SPAN, and PIQ.

4. General discussion

The relationship between EFs and conceptual development has been well established in the ToM literature and more recently in other domains, such as naïve biology (Zaitchik et al., 2013) and naïve physics (Baker et al., 2011). The present results replicate and extend the findings in the naïve physics literature. Similar to Baker et al. (2011), we found a trend toward a significant correlation (controlling for age, gender, and receptive vocabulary) between a delay inhibitory control composite variable and pretest performance on the tubes task. One possible interpretation of this result is that children who fail to inhibit the urge to find the ball as fast as possible ignore the tubes altogether and search directly below the place where they have seen the ball disappear. Furthermore, like Baker et al. (2011), we did not find a relationship between pretest performance on the tubes task and RC-EF, even though unlike Baker et al. (2011) we used a composite RC-EF variable. In the present study, however, we also observed a trend toward a statistically significant partial correlation between digit span and pretest performance on the tubes task and statistically significant bivariate and partial correlation between pretest performance on the tubes task and performance IQ, demonstrating a broader link between children's EF skills and their utilization of their physical knowledge.

The correlation between delay inhibitory control and performance on the tubes task has been taken as evidence for the expression alone account of developmental changes on the tubes task. On this view, the waning of the gravity error with age can be completely explained by children's maturing inhibitory control. That is, the correlation between EFs and pretest performance on the tubes task is observed because EFs are important for *expressing* the knowledge that is computed using physical reasoning processes available to infants, such as computations of spatiotemporal continuity of trajectories and knowledge of solidity. Difficulty expressing such knowledge could stem from a variety of sources, including conflict with simpler heuristic rules – like the expectation that objects will fall straight down – which often but not always match the outcome of reasoning based on a richer set of knowledge.

Contrary to the continuity hypothesis, the present data also confirm that verbal instruction concerning the role that the tubes play in guiding the movement of the ball improves children's performance. Repeated feedback from looking in the wrong place, and from being shown where the ball really is during the pretest trials, does not improve performance but, replicating previous work by Bascandziev and Harris (2010), the training intervention involving testimony about how the tubes constrain the movement of the ball does so. This fact demonstrates that children need to develop some explicit insight into how the tubes constrain the ball's trajectory, and it raises the possibility that EFs play a role not only in expressing that understanding, but also in constructing new explicit physical understanding. If so, we should expect to find a correlation between the rate of change on the tubes task and EFs.

The effects reported in the present study are admittedly small but consistent with the latter hypothesis. Controlling for age, gender, and receptive vocabulary, children who score higher on DSPAN and PIQ (and to some extent on RC-EF), improve more on the tubes task compared to children who score lower on those measures. This evidence supports the claim that EFs are important for the process of *constructing* knowledge that is necessary for solving the tubes task.

The design and the findings of the present study are in many ways parallel to the design and findings of Benson et al.'s (2013) study that examined the relationship between individual differences in EFs and the development of a Representational Theory of Mind. The results of that study were taken as evidence that RC-EF plays an important role in the process of *constructing* a Representational Theory of Mind, and not merely expressing ToM knowledge the child already has. The present results generalize this pattern of findings, and this overall argument, to a domain of intuitive physics.

However, one might possibly account for the correlation between EFs and rate of change on ToM knowledge or rate of change on the tubes task without implicating EFs in the construction of domain specific knowledge. That is, the findings of the present study and the findings of Benson et al. (2013) are open to the following alternative interpretation: both children with high EFs and children with low EFs might have benefitted from instruction, but only children with higher EFs were able to *express* the newly acquired knowledge. This account does not invoke any role for EFs in *constructing* new knowledge and yet it is consistent with the finding that individual differences in EFs are associated with the rate of change in performance. We see nothing in the Benson et al. (2013) data that militate against this alternative. Indeed, that children benefit from a single episode of feedback, and that children's improvement is limited to only that aspect of ToM that was the direct target of instruction, argue against the proposal that the feedback in the Benson et al. (2013) finding is sufficient for the construction of the explicit representational theory of mind achieved by 4-year-olds. Similarly, the effect of the single verbal explanation in the present study also argues against a major constructive change being achieved during training. Nonetheless, our data argue against the hypothesis that the role of high EFs is in the expression of the knowledge change achieved by our training rather than in the achievement of that knowledge change. This is why: different aspects of EFs were associated with pretest performance from those that were associated with degree of change as a function of training. For example, whereas delay inhibitory control was correlated with pretest performance in the present study and also in Baker et al.'s (2011) study, it was not correlated with the rate of change. Conversely, whereas RC-EF was not a good predictor of pretest performance both in the present study and in Baker et al. (2011), it was trending toward a significant predictor of the rate of change on the tubes task. If the only way that EFs affect performance on this task is through expression of knowledge, then we should have found that the same predictors are good predictors of both pretest performance and the rate of change. This aspect of our results supports the conclusion that EFs play an important role in learning and not only in the expression of what has already been learned.

4.1. Direction of influence

One possible interpretation of the present data is that the training influenced children's EFs, which in turn influenced their ability to *express* their already present knowledge about how the tubes constrain the movement of the ball. This is very unlikely. The training that children received in this study was very short and it was designed to target children's understanding of the role that the tubes play in guiding the trajectory of the ball (see script in the Method section). Thus, it seems unlikely that these simple verbal instructions, lasting approximately 2 min, influenced children's working memory, their RC-EF or their performance IQ. The more likely interpretation is that over the course of this experiment children's understanding of the role that the tubes play in guiding the trajectory of the ball changed whereas their EFs remained unchanged.

4.2. The role of performance IQ

Studies investigating the role of EFs in conceptual development often use measures of receptive vocabulary to control for verbal intelligence (e.g., see Baker et al., 2011; Benson et al., 2013; Carlson & Moses, 2001; Zaitchik et al., 2013). In the present study, we also included a receptive vocabulary task, which is a measure of crystallized IQ, to control for verbal ability, and we found that the correlations between EFs and baseline success on the tubes task survived after controlling for receptive vocabulary. This was also true of the correlations between EFs and improvement due to teaching, as Benson et al. (2013) found in the case of Theory of Mind. In addition, we included the block design and object assembly subtests of WPPSI-III, which are measures of fluid IQ. In contrast to our measure of crystallized IQ, which predicted neither pretest performance nor the rate of change on the tubes task, the performance IQ composite variable predicted both pretest performance and the rate of change when age, gender, and receptive vocabulary were controlled for.

There are at least two ways of thinking about the relationship between PIQ and performance on the tubes task. First, PIQ was correlated with all of the EF measures in our battery (RC-EF, D-EF, DSPAN, and SOP, although the correlation with SOP did not survive controlling for receptive vocabulary, age, and gender). Thus, it is possible that PIQ predicts both initial tubes performance and rate of change because of the contribution of EFs to PIQ. That is, PIQ can be seen partly as a complex cognitive control construct, built from elementary EFs. The other possibility is that PIQ has a unique predictive power. That is, although many studies have found that various EFs and especially working memory are correlated with fluid IQ, the fluid IQ variance is not exhausted by EFs (e.g., Friedman et al., 2006). Thus, it is plausible that the fluid IQ variance that is not associated by EFs is predictive of initial tubes performance and rate of change. To test this possibility, we fitted a multilevel model to data, which included PIQ, RC-EF, and DSPAN as predictor variables and Age, Gender, and Receptive Vocabulary as control variables. Controlling for age, gender, receptive vocabulary, RC-EF, and PIQ there is a trend such that there is gain of 0.07 on the tubes task for every 1-unit difference in DSPAN. In addition, controlling for age, gender, receptive vocabulary, RC-EF, and DSPAN there is a trend such that there is a performance gain of 0.08 on the tubes task for every 1-unit difference in PIQ. Thus, despite the moderate correlation between PIQ and DSPAN that can produce multicollinearity issues, the unique variance in PIQ after controlling for DSPAN and RC-EF and the unique variance in DSPAN after controlling for PIQ and RC-EF continues to be (albeit weakly) associated with the rate of change on the tubes task. The present findings suggest that measures of fluid IQ, in addition to measures of crystallized IQ, should also be included in studies relating EFs to conceptual development. Future research should explore in more depth the relationships between EFs and fluid IQ in young children.

4.3. In what ways could EFs influence children's understanding of the tubes mechanism?

In what way would children's developing understanding of the tubes mechanism depend on EFs? One very basic answer to this question is that children with higher EFs are capable of paying more attention to relevant information and are also capable of retaining that information. By implication, children with higher EFs would have the relevant information at their disposal for further processing, whereas children with lower EFs would not. Thus, it seems very likely that individual variance on these very basic processes would predict individual differences in the rate of change on the tubes task and also the rate of change on other processes that involve the construction of new knowledge.

Another possibility is that children with high EFs are better at drawing relevant inferences from instruction. Recall that children received an instruction saying that the ball could not escape from the tube and that it rolled inside the tube. In other words, the experimenter did not provide any explicit guidance about how to find the ball. Thus, it seems unlikely that remembering the particular instruction that children received in this study – without understanding what it meant for the movement of the ball inside the tube – could have helped them to improve. These considerations raise the possibility that EFs have helped children understand the implications of the verbal instruction. There are several reasons why this speculation seems plausible: a) in order to start contemplating new hypotheses, children need to be able to inhibit prepotent hypotheses (e.g., inhibit the hypothesis that objects always fall in a straight vertical line), and b) shift their attention between different pieces of information (e.g., away from the prepotent hypothesis and toward other factors that might be guiding the trajectory of the ball). Indeed, it would be impossible to even begin contemplating new hypotheses without having these abilities. For example, without inhibitory control children would indefinitely remain in the grip of the prepotent hypothesis (see Carlson & Moses, 2001; Moses & Tahiroglu, 2010 for a similar argument about the role of EFs in the development of Theory of Mind). Furthermore, without set shifting, children would be unable to suspend the prepotent hypothesis and flexibly shift their

mode of thinking towards generating a new hypothesis. Finally, without working memory, children would be unable to hold and manipulate the relevant information, which means that they would not be able to revise an existing prepotent hypothesis. Thus, by implication, children who have better inhibitory control, set shifting, and working memory ability, will be better equipped for *constructing* new hypotheses and new knowledge, in the face of cognitive conflict or new information.

4.4. Limitations and future directions

The findings of the present study are important because they show that there is an association between children's EFs and their ability to construct new explanatory understanding that helps them improve their performance on the tubes task. To our knowledge, the study by Benson et al. (2013) study is the only other recent study that has shown a similar relationship between EFs and children's improvement on a conceptually demanding task, in their case a theory-of-mind task. Taken together, the findings of these two studies demonstrate that EFs are associated with the construction of new, domain specific knowledge, and do so across two very different domains. This in turn raises the question of whether EFs are associated with the development of naïve framework theories across other domains, such as naïve biology and naïve physics. Future research might explore these questions.

One important limitation that these studies have in common is that they are correlational. That is, the nature of the results precludes us from ruling out third variable explanations. Therefore, future research might employ longitudinal experimental designs. More specifically, some recent studies have shown that a variety of interventions can influence the development of EFs (Diamond & Lee, 2011). Randomized controlled studies in which EFs are experimentally manipulated by training could be used to test the hypothesis that there is a causal relationship between EFs and the development of framework theories.

Another limitation of the present study is that the sample of children who participated in the present study scored approximately 0.6 standard deviations above the national norm on the Performance IQ measures and 1.1 standard deviations above the national norm on Receptive Vocabulary (see Table 1). Future research should explore whether the effects reported here would hold in samples with average and below average scores on normed IQ scales.

4.5. Conclusion

In conclusion, the findings of this study suggest that EFs mediate the construction of new explanatory understanding that helps children improve their performance on the tubes task. Future research should ask further questions about the relationship between EFs and the construction of new domain specific knowledge, scientific theories, and scientific concepts. The answers to these questions can help us gain a better understanding of the mechanisms that make the existence of rich conceptual knowledge in humans possible and it can also help us design better educational interventions.

Acknowledgments

This research has been supported by a grant from the National Science Foundation NSF INSPIRE BCS 1247396. We are grateful to the children and their families for participating in this study. We thank Terrence Tivnan for his comments on an earlier draft and Michelle Bang, Kristiana Laugen, and Maria Renken for help in data collection.

Appendix A.

General description of the multilevel models. We first describe the level 1 and level 2 submodels and we explain what each submodel represents. The level 1 submodel describes the shape of the trajectory of change. Given that there are only two repeated measures of our dependent variable (performance on the tubes task) we can only postulate a linear trajectory of change. Thus, the mathematical representation of the level 1 submodel is:

$$Y_{ij} = \pi_{0i} + \pi_{1i} TEST_{ij} + \varepsilon_{ij}$$

where TEST can take two values (0 for pretest and 1 for posttest). The dependent variable Y is performance on the tubes task for individual i at occasion j (there are two occasions: pretest and posttest). The Intercept (parameter π_{0i}) is the score on pretest (when TEST = 0) on the tubes task for individual i in the population. The parameter π_{1i} is the change from pretest to posttest on the tubes task for individual i in the population. The error ε_{ij} represents the residual variance across the two measurement occasions for individual i in the population (Singer & Willett, 2003).

The level 1 submodel specifies that children can have different pretest scores on the tubes task (π_{0i}) and different slopes (change from pretest to posttest performance on the tubes task (π_{1i})). However, the level 1 model does not say anything about the relationship between children's pretest scores and slopes on the tubes task and other predictor variables. The level 2 submodel does exactly that: it specifies the relationship between the initial status and the rate of change on the tubes task and other predictor variables (Singer & Willett, 2003).

Our research question is whether individual differences in EFs and PIQ are related to the rate of change on the tubes task. Thus, we can specify the level 2 submodel where the pretest performance on the tubes task (π_{0i}) and the change from pretest to posttest performance on the tubes task (π_{1i}) are the dependent variables and EFs and PIQ are the predictor variables. For

the sake of illustration, here we present the mathematical representation of the level 2 submodel with EF as a predictor variable only.

$$\pi_{0i} = \gamma_{00} + \gamma_{01} EF_i + \zeta_{0i}$$

$$\pi_{1i} = \gamma_{10} + \gamma_{11} EF_i + \zeta_{1i}$$

The parameter γ_{00} represents the population average score on pretest on the tubes task for children who have a score of 0 on EF. The parameter γ_{01} represents the difference in initial (pretest) scores on the tubes task for 1 unit difference in EF. The parameter γ_{10} represents the population average change from pretest to posttest on the tubes task for children who score 0 on the EF measures. The parameter γ_{11} represents the difference in population average rate of change for 1 unit difference in EF scores. This parameter (γ_{11}) addresses the research question of the present study, namely whether the difference in EFs (or PIQ) can predict a difference in the rate of change in children's performance on the tubes task. Finally, the terms ζ_{0i} and ζ_{1i} represent the population random variance in both initial status (pretest score) and slope (change from pretest to posttest).

Fitting controlled multilevel models for change to data. Given that we have created several different predictor variables for EFs, we fitted separate models to the data (one model for each predictor variable). Each of these models controls for age, gender, and receptive vocabulary. We first call attention to what the number 0 means across all variables included in the models. A score of 0 on EF and PIQ measures means average performance. In addition, in order to have interpretable intercepts, we centered the variable age on its mean by subtracting the mean from each individual value. Thus, the number 0 represents the average age in this sample. Finally, for the variable gender, the number 0 represents girls.

Table A1 presents the full controlled multilevel models for change. The parameter estimates presented in the upper panel of Table A1 (γ_{00} through γ_{04}) reflect the relationship between the predictor (and control) variables and the pretest performance on the tubes task. These parameter estimates should be consistent with the partial correlations presented above. Specifically, the interpretation of the parameter estimate γ_{00} (across all models) is that the predicted average proportion of correct searches on the tubes task for girls with an average age, average receptive vocabulary, and an average RC-EF, DSPAN, and PIQ is 0.37, 0.34, and 0.37 respectively. We next turn to the parameter estimate γ_{01}, which is statistically significant only for PIQ. The interpretation of this parameter estimate is that controlling for age, gender, and receptive vocabulary, on average, the predicted score on pretest on the tubes task is 0.11 higher for every 1-unit difference in PIQ. This result is consistent with the partial correlations presented below diagonal in Table 2. The parameter estimate γ_{02} (Age) was trending toward being statistically significant in the model where RC-EF was the predictor of interest and it was statistically significant in the model where D-SPAN was the predictor of interest. The interpretation of this parameter is that controlling for gender, receptive vocabulary, and RC −EF (or DSPAN), there was a trend such that on average, the predicted score at pretest on the tubes task was 0.04 higher for every 1-unit difference in age. However, there was no such effect of age when PIQ was held constant. Finally, the parameter estimate γ_{03} (Gender) means that controlling for age, receptive vocabulary and for RC-EF, DSPAN, and PIQ (in the three separate models), on average, the predicted score of boys at pretest was respectively 0.26, 0.32, and 0.25 higher than that of girls. These results are consistent with our correlational analyses with one notable exception. Namely, whereas the partial correlation between pretest performance on the tubes task and DSPAN was trending toward being statistically significant ($r = 0.21$, $p = 0.09$), the relationship between pretest performance on the tubes task and DSPAN was not statistically significant in the multilevel model (γ_{01} =0.08, $p = 0.12$). Similarly, the p value of the partial correlation between pretest and D-EF was $p = 0.08$ and the p value associated with the D-EF parameter estimate (pretest) in the multilevel model was $p = 0.12$. We believe that this small difference in p values between the partial correlations and the multilevel models is because the multilevel models are estimating many more parameters (initial status + rate of change) compared to the partial correlations (initial status only) and they are using a different method (maximum likelihood estimation). The small differences in p values was consistent across all models. Here we give the p values of the partial correlations and the p values of the multilevel initial status parameter estimates: Partial correlation between pretest and RC-EF p = 0.21, multilevel model RC-EF p = 0.29. Partial correlation between pretest and SOP p = 0.14, multilevel model SOP p = 0.18. Finally, partial correlation between pretest and PIQ p = 0.01, multilevel model PIQ p = 0.02. Thus, despite the differences in methodology and number of parameters estimated, the multilevel models and the correlational analyses produced very consistent findings, notwithstanding the small differences in p values.

The rate of change parameter estimates (lower panel of Table A1) are presented and discussed in the Results section of the paper.

Table A1
Parameter estimates and goodness-of-fit statistics of controlled multilevel models for change.

		Parameter	Predictor **RC-EF**	Predictor **DSPAN**	Predictor **PIQ**
Pretest performance	Intercept	γ_{00}	**0.37***** **(0.06)**	**0.34***** **(0.07)**	**0.37***** **(0.06)**
	Predictor Variable	γ_{01}	0.05 (0.05)	0.08 (0.05)	**0.11*** **(0.05)**
	Control Variable Age-Centered	γ_{02}	**0.04**$^{\sim}$ **(0.02)**	**0.04*** **(0.02)**	0.02 (0.02)
	Control Variable Gender	γ_{03}	**0.26**** **(0.10)**	**0.32**** **(0.10)**	**0.25**** **(0.09)**
	Control Variable Receptive Vocabulary	γ_{04}	−0.02 (0.05)	−0.0001 (0.05)	−0.04 (0.04)
Rate of change	Intercept	γ_{10}	0.07 (0.06)	0.04 (0.05)	0.06 (0.06)
	Predictor Variable	γ_{11}	**0.08**$^{\sim}$ **(0.04)**	**0.09*** **(0.04)**	**0.12*** **(0.05)**
	Control Variable Age	γ_{12}	−0.0006 (0.02)	**0.03**$^{\sim}$ **(0.02)**	0.009 (0.02)
	Control Variable Gender	γ_{13}	**0.15**$^{\sim}$ **(0.08)**	0.08 (0.09)	0.14 (0.09)
	Control Variable Receptive Vocabulary	γ_{14}	0.003 (0.04)	0.0001 (0.04)	−0.04 (0.04)
Variance Components					
Level 1	Within person	σ^2_ϵ	0.07*** (0.01)	0.06*** (0.01)	0.07*** (0.01)
Level 2	In initial status	σ^2_0	0.12*** (0.02)	0.12*** (0.03)	0.08*** (0.02)
	In rate of change	σ^2_1	–	–	–*
Goodness of fit	Deviance		154.4	106.3	133.1
	AIC		178.4	130.3	157.1
	BIC		207.5	156.8	186.3

Note: Standard errors in parentheses. The analyses were conducted with arcsine transformed dependent variable. The parameter estimates presented here are de-transformed.
*Bottom of Table presents the variance components and goodness of fit statistics. The variance components and associated statistical tests represent the remaining unexplained residual variability in outcome variable that could be explained by other predictors.
**In an effort to address "final hessian not positive definite," we have simplified the error covariance structure by eliminating the level-2 residual for rate of change and its associated covariance parameter. Key: $\sim p \leq .10$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

References

Appleton, M., & Reddy, V. (1996). Teaching three year-olds to pass false belief tests: a conversational approach. *Social Development*, *5*, 275–291.
Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423.
Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation*. Academic Press.
Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in 5-month- old infants. *Cognition*, *20*, 191–208.
Baker, S. T., Gjersoe, N. L., Sibielska-Woch, K., Leslie, A. M., & Hood, B. M. (2011). Inhibitory control interacts with core knowledge in toddlers'manual search for an occluded object. *Developmental Science*, 270–279.
Bascandziev, I., & Harris, P. L. (2010). The role of testimony in young children's solution to a gravity-driven invisible displacement task. *Cognitive Development*, *25*, 233–246.
Bascandziev, I., & Harris, P. L. (2011). Gravity is not the only ruler for falling events: young children stop making the gravity error after receiving additional perceptual information about the tubes mechanism. *Journal of Experimental Child Psychology*, *109*, 468–477.
Benson, J. E., Sabbagh, M. A., Carlson, S. M., & Zelazo, P. D. (2013). Individual differences in executive functioning predict preschoolers' improvement from theory-of-mind training. *Developmental Psychology*, *49*, 1615–1627.
Berthier, N. E., DeBlois, S., Poirier, C. R., Novak, M. A., & Clifton, R. K. (2000). Where's the ball? Two- and three-year-olds reason about unseen events. *Developmental Psychology*, *36*, 394–401.
Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function: and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*, 647–663.
Burgess, G. C., Braver, T. S., Conway, A. R. A., & Gray, J. R. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General*, *140*, 674–692.
Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.
Carlson, S., & Moses, L. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*, 1032–1053.
Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: Stability and prediction from ages 2 to 3. *Developmental Psychology*, *40*, 1105–1122.
Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*, 547–552.
Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: a meta-analysis. *Child Development*, *85*, 1777–1794.
Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.
Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, *318*, 1387–1388.
Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4–12 years old. *Science*, *333*, 959–964.
Duncan, J., Burgess, P., & Emslie, H. (1995). Fluid intelligence after frontal lobe lesions. *Neuropsychologia*, *33*, 261–268.
Engel de Abreu, P. M. J., Conway, A. R. A., & Gathercole, S. E. (2010). Working memory and fluid intelligence in young children. *Intelligence*, *18*, 346–356.

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., & Flavell, E. R. (1986). Development of knowledge about the appearance-reality distinction. *Monographs of the Society for Research in Child Development*, *51*(1, Serial No. 212)
Flavell, J. H., Green, F., & 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.
Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, *17*, 172–179.
Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory: and fluid intelligence in children. *Biological Psychology*, *54*, 1–34.
Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, *10*, 483–527.
Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2011). Quantity: not quality: the relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, *7*, 531–536.
Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*, 177–190.
Gathercole, S. E., Tiffany, C., Briscoe, J., Thorn, A. S. C., & ALSPAC Team. (2005). Developmental consequences of poor phonological short-term memory function in childhood: a longitudinal study. *Journal of Child Psychology and Psychiatry*, *46*, 598–611.
Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Sciences*, *7*, 287–292.
Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and ion: performance of children 3.5–7 years old on a Stroop-like day-night test. *Cognition*, *53*, 129–153.
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., & Graff, P. (1988). Knowing how you know: young children's ability to identify and remember the sources of their beliefs. *Child Development*, *59*, 1366–1371.
Hongwanishkul, D., Happaney, K. R., Lee, W. S., & Zelazo, P. D. (2005). Assessment of hot and cool executive function in young children: age-related changes and individual differences. *Developmental Neuropsychology*, *28*, 617–644.
Hood, B. M. (1995). Gravity rules for 2- to 4-year olds? *Cognitive Development*, *10*, 577–598.
Hood, B. M., Santos, L., & Fieselman, S. (2000). Two-year olds' naïve predictions for horizontal trajectories. *Developmental Science*, *3*, 328–332.
Hood, B. M., Carey, S., & Prasada, S. (2000). Predicting the outcomes of physical events: two-year-olds fail to reveal knowledge of solidity and support. *Child Development*, *71*, 1540–1554.
Hood, B. M., Wilson, A., & Dyson, S. (2006). The effect of divided attention on inhibiting the gravity error. *Developmental Science*, *9*, 303–308.
Hughes, C. (1998). Finding your marbles: does preschoolers' strategic behavior predict later understanding of mind? *Developmental Psychology*, *34*, 1326–1339.
Hughes, C., & Ensor, R. (2007). Executive function and theory of mind: predictive relations from ages 2–4. *Developmental Psychology*, *43*, 1447–1459.
Joh, A. S., Jaswal, V. K., & Keen, R. (2011). Imagining a way out of the gravity bias: preschoolers can visualize the solution to a spatial problem. *Child Development*, *82*, 744–750.
Kharitonova, M., & Munakata, Y. (2011). The role of representations in executive function: investigating a developmental link between flexibility and abstraction. *Frontiers in Psychology*, *2*, 1–12.
Kochanska, G., Murray, K., 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.
Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–1834.
Lee, V., & Kuhlmeirer, V. A. (2013). Young children show a dissociation on looking and pointing behavior in falling events. *Cognitive Development*, *28*, 21–30.
Leslie, A. M. (1982). The perception of causality in infants. *Perception*, *11*, 173–186.
Leslie, A. M., & Keeble, S. (1987). Do six-month old infants perceive causality? *Cognition*, *25*, 265–288.
Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, *105*, 489–512.
Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science*, *12*, 142–149.
Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex frontal lobe tasks: a latent variable analysis. *Cognitive Psychology*, *41*, 49–100.
Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development*, *82*, 661–673.
Morrison, F. J., Ponitz, C. C., & McClelland, M. M. (2010). Self-regulation and academic achievement in the transition to school. In S. D. Bell, & M. Bell (Eds.), *Child development at the intersection of emotion and cognition* (pp. 203–324). Washington DC: Am. Psychol. Assoc.
Moses, L. J. (2001). Executive accounts of Theory-of-Mind development. *Child Development*, *72*, 688–690.
Moses, L. J., & Tahiroglu, D. (2010). Clarifying the relation between executive function and children's theory of mind. In B. Sokol, U. Muller, J. Carpendale, A. Young, & G. Iarocci (Eds.), *Self- and social-Regulation: exploring the relations between social interaction, social understanding, and the development of executive functions*. New York: Oxford University Press.
O'Neill, D. K., & Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology*, *27*, 390–397.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs. *Science*, *308*(5719), 255–258.
Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: you catch it from your sibs. *Child Development*, *65*, 1228–1238.
Powell, L. J. & Carey, S. (2016). Executive function depletion in children and its impact on Theory of Mind. Manuscript submitted for publication.
Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Development*, *73*, 734–751.
Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind. *Psychological Science*, *17*, 74–81.
Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: modeling change and event occurrence*. New York: Oxford University Press.
Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: training children to understand belief. *Child Development*, *67*, 2967–2988.
Snyder, H. R., & Munakata, Y. (2010). Becoming self-directed: abstract representations support endogenous flexibility in children. *Cognition*, *116*, 155–167.
Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*, 587–592.
Spelke, E. S., Keestenbaum, R., Simons, D. J., & Wein, D. (1995). Spatiotemporal continuity: smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, *13*, 1–30.
Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*, 580–586.
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.
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*, 1–34.
Zaitchik, D., Iqbal, Y., & Carey, S. (2013). The effect of executive function on biological reasoning in young children: an individual differences study. *Child Development*, *85*, 160–175.
Zelazo, P. D., & Frye, D. (1997). Cognitive complexity and control: a theory of the development of deliberate reasoning and intentional action. In M. Stamenov (Ed.), *Language structure, discourse, and the access to consciousness* (pp. 113–153). Amsterdam: John Benjamins.