Published in final edited form as:

J Cogn Dev. 2011 January 1; 12(2): 169–193. doi:10.1080/15248372.2011.563485.

Executive Function in Preschool Children: Test-Retest Reliability

Danielle M. Beck,
Simpson University

Catherine Schaefer,
University of Minnesota

Karen Pang, and
University of Washington

Stephanie M. Carlson
University of Minnesota

Abstract

Research suggests that executive function (EF) may distinguish between children who are well- or ill-prepared for kindergarten, however, little is known about the test-retest reliability of measures of EF for children. We aimed to establish a battery of EF measures that are sensitive to both development and individual differences across the preschool period using Conflict and Delay subtests that had a cool (abstract) or hot (extrinsic reward) focus. Results from 151 children in three age groups (2.5, 3.5, and 4.5) suggested acceptable same-day test-retest reliability on all but Delay-Cool subtasks. These findings will inform appropriate measurement selection and development for future studies.

Executive functioning (EF) refers to higher-order, self-regulatory cognitive processes that aid in the monitoring and control of thought and action (for a review, see Zelazo, Carlson, & Kesek, 2008). Components of EF that have been identified include working memory, attentional flexibility, and inhibitory control processes including resistance to temptation. EF is primarily associated with the prefrontal cortex and most knowledge about EF stems from neuroimaging data and clinical observations of adults with brain lesions or medical conditions known to affect prefrontal functioning.

Several studies utilizing various measures of inhibitory control in children have suggested that EF may play an important role in cognitive and social development, such as theory of mind, pretend play, social competence, moral conduct, and school readiness (e.g., Blair & Razza, 2007; Carlson, Mandell, & Williams, 2004; Carlson & Wang, 2007; Shoda, Mischel, & Peake, 1990). Conversely, disruptions in EF are implicated in a number of childhood disorders and are associated with poor social and academic adjustment (Casey, Tottenham, & Fossella, 2002; Hughes, 1998; Rodriguez, Mischel, & Shoda, 1989).

Despite its significance in clinical, educational, and research settings, progress in the field of early EF development and its role in child outcomes has been hampered by a lack of appropriate behavioral measures of EF. Researchers in recent years have used a mix of table-top behavioral tasks, computerized tasks, tasks tapping multiple (but often unspecified) aspects of EF, tasks that are not developmentally sensitive in the age group of interest (i.e., they produce floor or ceiling effects), or tasks that are not child-friendly (which have a high

refusal rate). Furthermore, researchers often use the same measures over time to investigate development, however, very few of these tasks have been systematically evaluated for reliability. Each of these factors introduces measurement error in the construct of EF. The goal of this study was to assess test-retest reliability of a new battery of behavioral measures of EF.

Test-retest Reliability

Many EF tasks, such as Delay of Gratification, are only novel once, so it is difficult to establish reliability. Despite this, test-retest reliability has been investigated on a variety of measures where task novelty is not believed to relate directly to performance. Many of these studies, however, have used adult participants (e.g., Fan, McCandliss, Sommer, Raz, & Posner, 2002; Lowe & Rabbitt, 1998) or clinical populations, such as patients with sleep apnea or brain lesions (Ingram, Greve, Ingram, & Soukup, 1999; Tate, Perdices, & Maggiotto, 1998).

Research investigating test-retest reliability of EF measures in children has been limited and results have varied. Gnys and Willis (1991) reported test-retest reliability on the Tower of Hanoi to be very good ($r = .72$) with a 25-min delay between administrations. Although this value is fairly high in terms of conventional psychometric criteria (Coolican, 1994), mean scores on the second administration of the task were more than one standard deviation higher than on the first administration. Similarly, Aman, Roberts, and Pennington (1998) found significant improvement on the Tower of Hanoi retest among boys with Attention Deficit Hyperactivity Disorder (ADHD) and a control group, suggesting this task may be prone to practice effects. Bishop, Aamodt-Leeper, Creswell, McGurk, and Skuse (2001) also investigated test-retest reliability on the Tower of Hanoi (modified version), with a 30-40 min delay between tests, and reported a reliability score of $r = .50$, which is considered fairly low by psychometric criteria (Coolican, 1994). It is possible that differences in reliability across studies could be due to the differences in ages of the children sampled: 5-year-olds in Gnys and Willis versus 7-to-10-year-olds in Bishop et al. and 7-to-15-year-olds in Aman et al. Within their sample, however, Bishop et al. did not find any significant age differences in reliability scores. Test-retest reliability on a variety of other EF measures has been reported to be within an acceptable range. An extensive study utilizing several tasks of EF with 7-to-12-year-olds, including Self-Ordered Pointing, Day/Night Stroop, and Fruit Stroop reported good reliability on all tasks ($rs = .76$ to $.93$) with an average retest interval of 4 months (Archibald & Kerns, 1999). A major limitation of this finding, however, is that test-retest data were collected on a small number of participants ($N = 18$).

Importantly, a limitation of the existing literature is that no published studies have investigated test-retest reliability of EF tasks in preschool children. The preschool period has been shown to be a time of significant advances in the development of EF (e.g., Carlson, 2005) and the focus of many recent investigations of EF in relation to brain, cognitive, academic, and social development (for a recent review, see Zelazo et al., 2008). It is crucial, therefore, to establish reliability of EF measures in this age group to better evaluate research findings and aid in task selection for future research. Another major challenge in research on EF is that it is difficult to find tasks that can be administered repeatedly to children at a variety of ages, and little is known about the appropriateness of repeating those measures to assess EF development over time. It is our hope that the information gathered in this study will help address this gap. Having reliable and valid measures of EF is essential when examining the stability of individual differences in EF, as well as for drawing inferences regarding developmental milestones and group differences (e.g., in children with autism or ADHD).

EF Dimensions

We designed a battery of four tasks that crosses two major dimensions that have been identified and examined in recent studies of EF in preschool children: conflict versus delay, and hot versus cool. Conflict tasks require children to initiate a goal-directed behavior in the face of conflicting stimulus properties, such as in a Stroop task, or when there is proactive interference from a previous response, as in task- or rule-switching. Delay tasks, on the other hand, require children to dampen or postpone a dominant response or resist temptation in the short term for a larger delayed reward, such as in a delay of gratification task. Some studies that have included a large battery of both types of measures in preschoolers reported that these aspects of self-regulation emerged as separate factors in principal components analyses (Bernier, Carlson, & Whipple, 2010; Carlson & Moses, 2001; Davis-Unger & Carlson, 2008; Kochanska, Murray, & Harlan, 2000), although they were correlated ($r_s \approx .50$). Although performance across conflict and delay EF show age-related improvement during the preschool period (Carlson, 2005), they have shown different patterns of relations to other constructs. For example, children's performance on conflict EF, but not delay EF, is significantly related to parenting quality (Bernier et al., 2010), to children's theory of mind performance (Carlson & Moses, 2001; Carlson, Moses, & Breton, 2002) and effective teaching (Davis-Unger & Carlson, 2008). In addition, bilingual children perform significantly better on Conflict EF than monolingual children; there is no similar advantage on delay tasks (Bialystok & Martin, 2004; Carlson & Meltzoff, 2008). Conversely, delay EF, but not conflict, is positively related to sleep regulation in toddlers (Bernier, Carlson, Bordeleau, & Carrier, in press) and has shown robust correlations with a host of positive cognitive and social outcomes. For example, preschool age children's ability to delay gratification is positively related to their scholastic performance and ability to cope with frustration and stress as adolescents (Shoda, Mischel, & Peake, 1990). The different task demands of conflict versus delay may tap separate neurological structures. Delay tasks, for instance, have been shown to provide assessment of orbitofrontal cortex (OFC) function, whereas conflict tasks draw upon both the OFC and prefrontal cortex (PFC) (Bechara, Damasio, & Reuven, 2007; Berlin, Rolls, & Kischka, 2004; Zelazo & Cunningham, 2007).

The second dimension concerns the affective-motivational context of the task, with some measures being relatively "cool" (emotionally neutral) and others being relatively "hot" (emotionally motivating, usually in the presence of appetitive extrinsic rewards) (Metcalf & Mischel, 1999; Zelazo, Qu, & Kesk, 2009). Traditionally, most research has focused on cool EF. Cool EF task performance is associated with the dorsolateral and ventrolateral prefrontal cortex (DL/VL-PFC) and is likely elicited by certain logic problems or tasks such as the Wisconsin Card Sort Task (WCST; Grant & Berg, 1948) or Dimensional Change Card Sort (DCCS; Frye, Zelazo, & Palfai, 1995; Zelazo, 2006) where cards must be sorted flexibly by various characteristics (e.g., shape and color). (See Moriguchi & Hiraki, 2009, for an imaging study implicating VL-PFC activation in young children during the DCCS.) Hot executive function, on the other hand, is typically correlated with activation in the orbitofrontal cortex (OFC), which has close connections with the limbic system (e.g., reversal learning of a rewarded response, Overman, 2004). Thus, situations that are emotionally charged or social in nature tend to call on hot executive functioning and OFC activation. It is important to note, however, that the "temperature" of any EF task is best characterized as a matter of degree, or the extent to which it features relatively implicit/intrinsic (cool) versus explicit/extrinsic (hot) motivation (Metcalf & Mischel, 1999).

Despite general improvements in EF across the preschool period, research has shown differential patterns of development for cool and hot EF as well as meaningful individual differences in task performance within a particular age range. However, relatively little is known about how motivational or affectively charged stimuli influence young children's EF. Under some circumstances, children seem to perform better on hot versions of EF tasks (Qu

& Zelazo, 2007; Zelazo et al., 2009), although other reports suggest children do better on cool versions (Carlson, Davis, & Leach, 2005; Prencipe & Zelazo, 2005). Similar to the distinction between conflict and delay tasks, individual differences on tasks of hot and cool EF have been shown to have distinct relations to concurrent and later functioning. For example, preschoolers who performed better on cool measures of EF were shown to have higher verbal mental age and academic achievement, whereas no such relations to hot EF were found (Brock, Rimm-Kaufman, Nathanson, & Grimm, 2009; Hongwanishkul, Happaney, Lee, & Zelazo, 2005).

Note that these two dimensions (conflict/delay and hot/cool) are not consistently found to warrant separate factors in the EF literature (e.g., Beck & Carlson, 2005; Carlson et al., 2004) and there is debate in the field as to which dimensions various EF tasks measure (Garon, Bryson, & Smith, 2008). However, it is important to note that the cognitive and affective dimensions of EF are proposed to be part of a single, interactive system (Metcalf & Mischel, 1999; Zelazo & Cunningham, 2007). Thus, it is difficult, if not impossible, to design a pure task that measures only one dimension. More often, tasks are designed to emphasize one dimension more than the other.

An additional complication is that these dimensions are often confounded in the EF literature, with conflict tasks being relatively cool and delay tasks being relatively hot. To address this problem in the current literature, we set out to cross these two dimensions of EF: conflict (inhibiting a prepotent response and initiating a conflicting one) and delay (postponing a response or outcome) with hot (motivational; extrinsic rewards) and cool (abstract; no direct extrinsic rewards) task properties. Our goal was to develop a smaller task battery than previous studies of EF have employed, yet one that is comprehensive and reliable. The first step in this research program, which will be discussed in this paper, was to establish the test-retest reliability of our newly adapted measures in children ages 2.5 to 4.5 years.

Method

Participants

Participants were 151 typically developing children ranging in age from 27 to 55 months ($M = 41.79$, $SD = 9.89$). They included 51 2.5 year-olds ($M = 30.00$ months, $SD = .87$, $range = 27$ to 33 months), 48 3.5 year-olds ($M = 41.33$ months, $SD = .81$, $range = 40$ to 43 months), and 52 4.5 year-olds ($M = 53.79$ months, $SD = .70$, $range = 52$ to 55 months). Forty-nine percent of the children were female, with an even distribution of boys and girls in each age group. The majority of participants (74.8%) were of a white/non-Hispanic ethnic background, 5.4% were white/Hispanic, 2% were Asian and 17.7% were bi-racial. Parents of four participants declined to state their ethnic background. Median family income was \$85,000 to \$100,000. Participants were recruited from flyers posted at local daycares and a university database of families who expressed an interest in being contacted about research studies. Data from 25 additional participants (18 2.5-year-olds, 11 3.5-year-olds, and two 4.5-year-olds) were not included in analyses because they did not complete all of the tasks due to shyness, fatigue, or non-compliance. An additional six children (four at age 2.5, one at age 3.5, and one at age 4.5) were not included in analyses because they were incorrect on the task rule checks.

Procedure

Participants were tested individually in a quiet laboratory playroom. Sessions were videotaped and lasted approximately 45-60 min. Children were randomly assigned (within age-group) to one of four conditions: Conflict-Hot, Conflict-Cool, Delay-Hot, or Delay-

Cool. Additionally, all children were given developmentally appropriate measures of working memory and intelligence. To evaluate the test-retest reliability of the measures of EF, children were administered tasks in the following fixed order: EF task (test), working memory task, EF task (retest), and abbreviated intelligence assessment. Delay between test and retest of EF tasks was approximately 15 min.

Measures

Working Memory

Spin the Pots (Hughes & Ensor, 2005): Children were shown a Lazy Susan with 8-12 visually distinct boxes and 6-10 distinct stickers, depending on their age: 8 boxes and 6 stickers at age 2.5; 10 boxes and 8 stickers at age 3.5; 12 boxes and 10 stickers at age 4.5. Children were then asked to place each sticker inside a box. The experimenter pointed out that there were not enough stickers for all the boxes and that two would remain empty. She then explained that each time the Lazy Susan was rotated, children could choose one box and see if there was a sticker inside, keeping the sticker if found and placing it on a sheet of paper. Next, the Lazy Susan was covered with an opaque scarf and rotated 360 degrees. The task ended when children found all hidden stickers or when the maximum number of spins was reached (12 spins maximum were given at 2.5 years, 16 at 3.5, and 20 at 4.5). Final scores were calculated as the proportion of stickers found to the total number of spins required to find all stickers (or the maximum number of spins allowed).

Intelligence

Stanford-Binet Intelligence Scales for Early Childhood (5th ed) (Roid, 2005): Two subtests from this assessment were used in this study: Verbal Knowledge (Vocabulary) and Nonverbal Fluid Reasoning (Object Series/Matrices). These two subtests provide an estimate of overall cognitive functioning (Abbreviated IQ; Roid, 2005). The Verbal Knowledge subtest measured children's productive language ability by requiring them to identify picture vocabulary and to define increasingly difficult words. The Nonverbal Fluid Reasoning subtest measured children's ability to solve novel figural problems and identify sequences of pictured objects or matrix-type figural and geometric patterns. Testing was discontinued when children responded incorrectly on four consecutive items. Abbreviated IQ scores (standard scores with a norm-referenced mean of 100 and $SD = 15$) were calculated from the two subtests.

Executive Function Measures

Conflict Scales: Administration of the Conflict EF task was designed to resemble that of intelligence tests. It consisted of four levels that shared a common core but increased in difficulty. The task began at a specific level given the age of the participant. Depending on children's success or failure at this level, the experimenter would proceed to a higher or more difficult level of the task (until a ceiling score was established) or recede to a lower or easier level of the task (until a basal score was established). The decision rules for task administration are illustrated in Figure 1.

The Conflict scale consisted of four levels: Categorization/Reverse Categorization, Dimensional Change Card Sort (DCCS)–Separated, DCCS–Integrated, and DCCS–Advanced. Each level represented a measure of Conflict EF and one of the simplest paradigms to assess task-switching performance. Children were required to attend to a relevant stimulus and categorize items accordingly. During the post-switch phase of each level, children must then inhibit attention to the particular feature of the stimulus that was no longer relevant and disinhibit attention to the feature that was previously ignored. Several

studies have shown correlations between the DCCS and independent measures of working memory and inhibition (Zelazo, Muller, Frye, & Marcovitch, 2003; Zelazo et al., 2008).

Level 1: Categorization/Reverse Categorization (Carlson et al., 2004): Children were presented with two small buckets with lids and two categories of stimuli (e.g., Teddy Graham and Goldfish) and asked to categorize stimuli accordingly by placing them in the corresponding buckets. Before beginning the task, the experimenter provided a demonstration and followed with a rule-check. Six test trials were then administered and the number of correct trials was recorded (6 max). If children were correct on at least five out of the six trials, the experimenter proceeded to Reverse Categorization (post-switch). Children were told to play a “silly” game and reverse the sorting scheme that was used in Categorization. The procedures for the Hot and Cool versions of Categorization and Reverse Categorization follow below.

Hot: Children were given a Goldfish cracker and a Teddy Graham to taste. They were then introduced to two buckets. One displayed a cartoon picture of a “mommy” Goldfish cracker and the other a picture of a “mommy” Teddy Graham. Children were told they would receive the Goldfish and Teddy Graham treats to take home with them if they followed the rules of the game. Before beginning the task, the experimenter demonstrated that the “baby” Teddy Graham treats would go in the bucket with the “mommy” Teddy Graham on it, and that the “baby” Goldfish would go in the bucket with the “mommy” Goldfish on it (one demonstration of each was given). A rule check followed for each: “Where does the baby Goldfish/ Teddy Graham go?” If children did not answer correctly, the rule was re-stated, and the question repeated. On each test trial, children were reminded of the rule (e.g., “Goldfish go here and Teddy Grahams go there.”) The experimenter then presented the stimuli (e.g., “Here’s a baby Goldfish”) and asked children to sort real Goldfish crackers and Teddy Grahams into their corresponding “mommy” buckets for six trials (e.g., “Where does it go?”) A total of three Goldfish and three Teddy Grahams were presented in the following order: G, TG, G, TG, TG, G).

If children were correct on at least five out of the six trials, the experimenter proceeded to Reverse Categorization. Children were told to play a “silly” game and reverse the sorting scheme that was used in Categorization, placing Goldfish crackers into the “Teddy Graham” bucket and Teddy Grahams into the “Goldfish” bucket for six trials in the order: G, TG, G, TG, TG, G.

Cool: Children were introduced to two buckets; one displayed a picture of a woman and the other a picture of a baby. They were then presented with large and small toy models of six different animals and told that they were “mommies” and “babies” of the respective animals (e.g., “Here’s a mommy elephant, and a baby elephant.”) Before beginning the task, the experimenter explained that the “baby” animals go in the bucket with the baby on it, and that the “mommy” animals go in the bucket with the mommy on it. Then, she demonstrated three trials of each. (Two more examples were given than in the Hot version because the exemplars varied across trials in this version of the task.) A rule check was administered as described above and test trials began. The experimenter reminded children of the rule on each trial (e.g., “Mommies go here and babies go there”) and then presented the animal (e.g., “Here’s a baby pig. Where does this one go?”). Children sorted baby and mommy animals for six trials in the order: B horse, M elephant, B elephant, M pig, M horse, and B pig.

In the Reverse Categorization phase, children were given similar instructions to those in the Hot version of the task and told to reverse the sorting scheme that was used in Categorization, placing the mommy animals into the “baby” bucket and the baby animals

into the “mommy” bucket for six trials in the order B tiger, M cow, B cow, M polar bear, M tiger, and B polar bear.

Level 2: Dimensional Change Card Sort – Separated (Diamond, Carlson, & Beck, 2005): Children were introduced to two black recipe boxes with slots cut in the top. Target cards were attached to the front of each box. The target cards consisted of a colored background and a black shape located on the center of the card (details below). Children were told to sort cards (in which the color/shape properties were opposite to those on the target cards) according to one dimension (e.g., color) for six trials and then to sort according to the other dimension (e.g., shape) for six trials. We followed the method of Zelazo (2006), in which the experimenter announced the rule before each trial, and presented a card and labeled it according to the current dimension (e.g., on a shape trial, “Here’s a boat. Where does it go?”).

Hot: Target cards consisted of a brown card (the color of Teddy Grahams) with a black silhouette of a Goldfish in the center and an orange card (the color of original Goldfish crackers) with a black silhouette of a Teddy Graham in the center. Sorting cards consisted of orange cards with a Goldfish silhouette and brown cards with a Teddy Graham silhouette. To make the task “hotter” with an incentive, children were told they would receive the Goldfish and Teddy Graham treats to take home with them if they followed the rules of the game.

Cool: Target cards consisted of a red card with a black silhouette of a truck in the center and a blue card with a black silhouette of a star in the center. Sorting cards were blue cards with a truck silhouette and red cards with a star silhouette. The procedure was identical to the Hot version, except there was no mention of rewards for good performance.

Level 3: Dimensional Change Card Sort - Integrated (Zelazo, 2006; Zelazo et al., 2003): This task followed the same procedure as DCCS-Separated except the stimuli contained a greater degree of perceptual conflict.

Hot: Target cards consisted of a white card with a brown Goldfish in the center and a white card with an orange Teddy Graham in the center. Children were instructed to sort cards (orange goldfish and brown teddy grahams) according to shape and then by color (six trials each). They were told they would receive the Goldfish and Teddy Graham treats to take home with them if they followed the rules of the game.

Cool: The procedure was identical to the Hot version except the target cards were white with red trucks or blue stars in the center, sorting cards were blue trucks and red stars, and rewards were not mentioned.

Level 4: Dimensional Change Card Sort - Advanced (Zelazo, 2006): A fourth sorting level was added in which some cards had a black border around the card and some did not. Otherwise, these cards were the same as the cards used in the integrated condition. The experimenter told children that if the card had a black border around it, they should play the “color game” (sort by color) and if the card did not have a black border, they should play the “shape game” (sort by shape). The experimenter performed one demonstration of each. In the Hot version, children were then reminded they could take the treats home if they played by the rules of the game. The advanced level consisted of 12 trials (6 border, 6 non-border, presented in the following order: B, NB, B, B, NB, NB, B, B, NB, NB, B, NB).

Scoring

Starting levels were the same for both test and retest for a given child. The starting level of the Conflict Scales depended on age: Categorization for 2.5-year-olds; Separated DCCS for 3.5-year-olds; and Integrated DCCS for 4.5-year-olds. If children passed a given level (5 or more correct out of 6 post-switch trials), then the experimenter continued to the next level. When children failed, however, the experimenter administered the next level downward as needed to establish a basal level. See Figure 1 for the Conflict Scale administration protocol. To compute a total Conflict EF score, the number of cards correctly sorted (out of a possible 48) was recorded. If children were not administered a particular level because their basal level was established above that level, then it was assumed that they would have correctly sorted the cards/items and the appropriate number of points were added to their total score. For example, if a 4.5-year-old scored 6 out of 6 correctly on Level 2 then Level 2 would be considered their basal level and they would receive 12 points for Level 1 (6 preswitch, 6 postswitch) even though they were not administered Level 1. If children failed a level, the task was stopped and they received no points for all subsequent levels.

Delay Tasks

Unlike Conflict, this task did not increase in difficulty or regress to an easier level. Rather, children received the same task regardless of age, although the instructions differed to ensure that even the youngest children understood the rules. Our Delay procedure was based on that of Thompson, Barresi, and Moore (1997) and Prencipe and Zelazo (2005). Children made a series of choices as to whether a reward should be received immediately or saved until “later, after we’ve finished all of our games for the day.” The rewards consisted of a favorite snack treat (e.g., Froot Loops or Goldfish crackers), stickers, and pennies presented in shallow clear plastic trays (2” × 2”). There were a total of nine trials, with three trials of each reward type, in a fixed order: 1 vs. 4 pennies; 1 vs. 2 stickers; 1 vs. 6 pennies; 1 vs. 4 stickers; 1 vs. 2 treats; 1 vs. 6 stickers; 1 vs. 4 treats; 1 vs. 2 pennies; 1 vs. 6 treats. On each trial, the experimenter gave children a choice between a smaller, immediate reward or a delayed, larger reward. Performance was evaluated according to the number of trials in which children chose to delay rewards. Children were given a score of zero if they chose the immediate reward or a one if they chose the delayed reward (max of 9). Two versions of the task were used, one with 4.5-year-olds and one with 2.5- and 3.5-year olds.

4.5-year-old Version

Children were told they would be playing a game with treats, stickers and pennies and they would be able to make choices about whether the rewards should be received now later, after all the games were completed. Three letter size envelopes labeled “treats,” “stickers,” and “pennies” were placed on the table along with a blank sheet of paper for stickers and a jar to hold pennies. The experimenter explained that treats which were chosen for later would be placed in the respective envelopes until after all the games were completed. If immediate choices were made, children were told that the pennies would go in the jar or in a pant pocket, snack treats would be eaten immediately and stickers would be placed on the blank sheet of paper. The experimenter gave two demonstrations, first choosing one Goldfish cracker to eat immediately (vs. four for later) and then saving two stickers for later (vs. one for now). Next, children were reminded of the rules and instructed on what would happen to the rewards depending on the choices they made. The experimenter then administered nine test trials. No feedback was provided during test trials.

Hot—Children were asked to make a series of choices for *themselves*. After the two demonstration trials, the experimenter explained, “Okay, now it’s your turn to choose. You’ll get to make some choices as to whether you want some stickers, pennies, or

[Goldfish crackers] for now or later, when we're all done playing our games for today... What do you want to do? Do you want to choose one penny now or four for later?"

Cool—Children were asked to make a series of choices for someone *else*, in this case, the experimenter. The experimenter explained that children would get to make choices as to what they thought *she* should do on each trial, (e.g., "In this game, you'll have the chance to make choices for me... For each choice you will need to choose what you think I should do. Should I choose to have the reward now, or wait until later after we've finished all of our games for the day ... What do you think I should do? Should I choose one penny now or four for later?")

2.5- and 3.5-year-old Version

To make the rules easier for younger children to understand and to keep children engaged, a familiar puppet (Elmo and/or Ernie) was added. The puppet was used to demonstrate the rules of the game to children so that they could see how specific choices and consequences were associated. As well, instead of three separate envelopes for each reward type (pennies, stickers, snack treats), 2.5- and 3.5-year-olds were given one 9" × 11" manila envelope with a picture of a house on the front in which they were to place their delayed rewards "for home." The house was intended to make it clear that the rewards inside the envelope were to be saved until the end of the game and could be taken home with them. The "for home" wording replaced "for later" because it was more concrete for younger children. In addition, the consequences of each choice were added to the instructions of each test trial. For example, for each trial the experimenter explained, "Do you want to choose one [treat to eat] right now or four to bring home?"

Hot—There were two demonstration trials in which Elmo (animated by the experimenter) chose a delayed reward (two treats for home versus one treat for now) and then an immediate reward (one penny for now versus six pennies for home) for himself. Elmo placed delayed rewards in a small envelope that was identical to the children's except for its size. It was explained that things that went in the envelope were to save for later, after all the games were finished and it was time to go home. After Elmo's choices were made, he demonstrated two trials in which he encouraged children to choose one immediate food reward (one treat for now versus four for home) and then delay a sticker reward (two stickers for home versus one for now). These demonstration trials were intended to allow children to practice making choices and associate the choice with the corresponding consequence. Next, Elmo said goodbye, took his "home" envelope with the treats inside and was put out of view. The experimenter then explained that it was the children's turn to make some choices and presented nine test trials (e.g., "Do you want to choose one penny to put in your cup right now or four to bring home?") in the same fixed order as described above.

Cool—In this condition, the experimenter first introduced a puppet (Ernie) to perform two demonstration trials in which she announced which choice she thought Ernie should make. The experimenter explained, "I get to make some choices for Ernie. I get to decide if Ernie should have some things for right now, or some things for later. If I decide that Ernie gets something for now, he can have it right away. But, if I decide he should have it for later, it goes in his home envelope so that he can have it later, when it's time to go home." The experimenter then performed two demonstration trials in which she announced which choices she thought Ernie should make (e.g., "I can choose for Ernie to have one treat right now, which means he can eat it right now, or I can choose for Ernie to have two treats to bring home. If I choose two for home, he can't have them right now; I have to put them in his Home envelope for when it's time to go home, later, after we're all done playing. This time, I choose one for now. So Ernie gets to eat it!") On the next demonstration trial the

experimenter made the delayed choice (four stickers for home versus one for now) for Ernie. At the conclusion of the demonstration trials, children were told that Ernie was done playing, and he would be taking his items home with him in his Home envelope. He was then placed out of view.

Next, the puppet Elmo was introduced to children. The experimenter asked children to decide what *Elmo* should do, just like the experimenter had decided what Ernie should do. Elmo was included because we thought the younger children would be more comfortable indicating what Elmo should do rather than what the experimenter should do, as was done for the 4.5 year-olds. The experimenter presented nine test trials in the fixed order described above (e.g., “Do you think Elmo should have one sticker to put on his paper right now or six to bring home?”) The procedure in its entirety (including the training phase) was repeated at retest.

Results

Preliminary Analyses

Means and standard deviations for the four conditions (Conflict-Hot, Conflict-Cool, Delay-Hot, and Delay-Cool) are presented in Table 1. Data were submitted to a series of one-way analyses of variance and results revealed no significant differences between conditions on age, sex, intelligence or family income level (all $ps > .10$) indicating no systematic differences occurred in group assignment.

First, a 2×3 between subjects factorial ANOVA was conducted on the data from the first administration of the tasks to investigate possible differences in scores on the Conflict scale as a function of motivational valence (Cool, Hot) and age group (2.5, 3.5, and 4.5). A significant main effect for age group was found, $F(2, 72) = 75.89, p < .001, \text{partial } \eta^2 = .68$. Tukey’s post-hoc tests indicated that children’s performance improved significantly in a stepwise fashion at each age group ($ps < .001$), and this was true for both the hot and cool versions of the Conflict scale considered separately. No significant main effect of motivational valence or interaction was found ($ps > .10$). Data from Delay tasks were submitted to the same statistical test and results revealed no significant main effects or interaction (all $ps > .10$). Results were similar when the same analyses were conducted using retest data for both Conflict and Delay.

Next, we investigated the relation of the EF tasks to working memory and intelligence. Conflict-Cool was significantly related to working memory (Spin the Pots task), $r(37) = .47, p < .01$, but significance was attenuated after controlling for age, $r(37) = .12, p > .10$. Conflict-Hot, Delay-Hot and Delay-Cool showed no significant relation to working memory ($rs = .08, -.04, \text{ and } .03, ps > .10$). None of the four EF tasks were significantly correlated with age-standardized IQ (Stanford-Binet), rs ranged from $-.23$ to $.25$, all $ps > .10$.

To further investigate the relation between age and each task, partial correlations were conducted controlling for working memory. Results showed a significant relation of age to both Conflict-Hot and Conflict-Cool ($rs = .84$ and $.89, ps < .001$). Results were not significant for Delay-Hot or Delay-Cool.

Test-Retest Analyses

For the main analyses, test-retest reliability was assessed. Figure 2 illustrates the relations between test and retest in each condition. Intraclass correlations ($ICCs$) were computed because this takes into account the shared method variance of identical tasks and is considered more appropriate and conservative than Pearson r for test-retest reliability analyses (McGraw & Wong, 1996). $ICCs$ were within or above accepted levels of 0.75 –

0.80 (Portney & Watkins, 1993) for all analyses, with the exception of Delay-Cool (see Figure 2). Partial correlations were also conducted using Abbreviated IQ scores as a covariate. Correlational results were similar to the ICCs (Conflict-Hot, $r = .83$; Conflict-Cool, $r = .94$; Delay-Hot, $r = .74$; Delay-Cool, $r = .52$; all $ps < .01$). Paired t -tests revealed no significant differences in means between test and retest, but Wilcoxon signed rank tests, which indicate whether the direction of change was systematic, revealed marginal differences in Conflict-Cool ($z = -1.81, p < .10$) and Delay-Hot ($z = -1.91, p < .10$), suggesting that children tended to do better on test than retest (i.e., mild fatigue effects). When Wilcoxon signed rank tests were conducted for each task according to age group, the only result approaching significance revealed 3.5-year-olds performed marginally better on test than retest on Conflict-Cool ($z = -1.90, p < .10$). There was no evidence of practice effects on any of the four tasks.

The distribution of scores for Conflict-Cool was investigated further as it appeared the distribution was bimodal (Figure 2b). Four outliers were removed because they did not fit the bimodal pattern of the distribution (two 2.5-year-olds and two 3.5-year-olds). Investigation of the two clusters revealed that although the cluster representing those with lower scores (0-20) showed adequate variability ($M = 11.57, SD = 5.93$ for test; $M = 9.62, SD = 4.65$ for retest), the cluster representing higher scores showed very little variability ($M = 42.07, SD = 1.49$ for test; $M = 41.73, SD = .80$ for retest). This evidence suggests a ceiling effect may have occurred. In fact, all 4.5-year-olds in this group ($n = 14$) scored between 40 and 46. Although 4.5-year-olds were not reaching the maximum score on this task ($max = 48$), their mean scores ($M = 41.93, SD = 1.44$ for test; $M = 41.71, SD = .82$ for retest) suggest that they were perseverating on one dimension (color or shape) during the DCCS-Advanced, thus scoring 50% (6 out of 12) on that subtask which would result in a total score of 42 out of 48. The ceiling effect may have attenuated or inflated test-retest reliability estimates for Conflict-Cool and so the ICC was re-computed with the 4.5-year-olds excluded (but including the four outliers). Results showed that reliability was lower than when the 4.5-year-olds were included but still very high ($ICC = .76$).

In the final set of analyses, test-retest reliability was investigated within each level of the Conflict scale (e.g., Categorization/Reverse Categorization, DCCS Separated, DCCS Integrated, and DCCS Advanced). ICCs for the Hot and Cool versions of the tasks are presented separately in Table 2. Subtasks within the cool version of the Conflict scale showed moderate to excellent reliability. Two subtasks that comprise Conflict-Hot were below the preferred level for reliability (Categorization/Reverse Categorization and DCCS Advanced), although the Hot versions of DCCS Separated and DCCS Integrated were highly reliable. It is difficult to speculate why the subtasks vary in their reliability. Separate analyses by age group did not explain the lower test-retest reliability found for these two levels, and descriptive statistics suggested that restricted range or variability were not to blame. Instead, it appears that the scales as a whole were highly reliable despite lower reliability of certain subtasks.

Discussion

Valid and reliable measurement tools are essential for the burgeoning field of theory and research (both basic and applied) concerning the development of “prefrontal” or executive functions in young children. A number of measures have been employed in recent research, each facing a variety of challenges including age-appropriateness and the blurring of dimensions such as conflict/delay and hot/cool. Surprisingly, none have been reported to have adequate retest reliability despite fairly wide use in the field, such as the DCCS. Hence, the primary aim of the present study was to examine test-retest reliability of a modified set of EF measures for children ages 2.5 to 4.5 years.

To address the age-appropriateness problem, we designed Conflict EF scales to trace the development of this aspect of EF using a set of tasks with a common core that increased in difficulty, as well as delay EF tasks that provided additional scaffolding for younger children. To address the confound that relatively “cool” (abstract) EF tasks for preschoolers tend to be conflict tasks, whereas relatively “hot” (reward-focused) EF tasks tend to involve delay, we added a Conflict-Hot task and a Delay-Cool task to our battery.

All tasks proved to be straightforward to administer from the examiner’s point of view and highly engaging from the child’s point of view. The Conflict scales succeeded in tapping the major maturational changes in EF across the preschool period, with significant advances from 2.5 to 3.5 to 4.5 years. The fact that even 4.5-year-olds did not reach the maximum score on the Advanced DCCS level suggests that the scales are useful up to at least age 5, and perhaps beyond when it is employed with a more broadly representative sample of children (e.g., a wider range of SES and parental education). Scores on the Hot and Cool versions – and their age trends – were nearly identical, despite the fact that the stimuli were necessarily different.

In contrast, we did not find age-related changes in performance on the Delay Choice task, unlike Prencipe and Zelazo (2005) and Thompson et al. (1997), who found 4-year-olds delayed more often than 3-year-olds in the “hot” condition only. It is possible that our modifications to the instructions and training with Elmo and his “home envelope” made the task more understandable even for 2.5-year-olds, who made just as many delay choices as 4.5-year-olds, on average, and delay scores were relatively low across all ages. In future research we will use uniform training procedures for all ages to see if developmental differences emerge. However, it is also notable that the standard deviations on Delay tasks were large at each age, and proportionally larger than on the Conflict scales, suggesting that individual differences (e.g., temperament) play a significant role in impulse-control, as measured by this task. Indeed, in Mischel’s classic delay of gratification paradigm, age differences are not often found below age 4 or 5 years. Instead, individual differences in delay time and the strategies employed to endure the delay were more important in terms of long-term prediction of later EF and child adjustment outcomes (Eigsti et al., 2006; Shoda et al. 1990). Hence, it is possible that conflict EF is more strongly based in age and brain maturation and that individual differences in the *rate* of acquisition are meaningful, whereas delay EF is a potentially informative individual-differences measure at any age. Future research should investigate the sources of individual differences that contribute to performance on delay EF.

To address the main question of this study, we found overall very good same-day test-retest reliability ($ICCs \geq .75$) on three of the tasks: Conflict-Cool, Conflict-Hot, and Delay-Hot. This finding is promising for future research utilizing these measures. It is especially significant for the Integrated-DCCS task we included as Level 3 in the Conflict scales, which has become the most pervasive measure of EF for 3- to 5-year-olds around the world (Zelazo, 2006). The test-retest reliability of the standard (cool) version of this task examined separately was $ICC = .94$.

Exceptions were found for the Delay-Cool task and the individual age-group analyses. On Delay-Cool, although $p < .01$, retest reliability did not meet psychometric standards ($ICC = .49$). Recall that on this task, children were asked to make a series of choices for now versus for later where someone else (the experimenter or a puppet) would receive the rewards. It could be argued that preschoolers were not very motivated to maximize the long-term rewards for someone else, however, prior research suggested that 3-year-olds delayed more often for other than for self, possibly because the “hot” reward focus had been defused (Prencipe & Zelazo, 2005). We did not find solid evidence for this advantage, although the

trend was for children to perform better (delaying more) on Delay-Cool than Delay-Hot on both test and retest. It is possible that we did not have enough power to detect a difference, as the only age group that was directly comparable to Prencipe and Zelazo's study (3.5 years) had a much smaller N in our study. It is notable also that the standard deviations were generally larger for Delay-Cool than Delay-Hot, especially on retest, suggesting there was more variability and perhaps the task of choosing for someone else seemed strange or frustrating for children to repeat. Indeed, most children in this condition asked, at some point, when it would be *their* turn to get some treats.

Importantly, and in contrast with previous research using the Tower of Hanoi (e.g., Aman et al., 1998), we found no significant practice effects on the EF tasks. This could be useful information for investigators wishing to administer the same measure more than once to examine the effects of development or a training intervention, as it would suggest any improvements in performance from Time 1 to Time 2 were not due merely to repeated exposure to the task. On the flip side, there was a marginally significant trend showing fatigue effects on Conflict-Cool and Delay-Hot. Children tended to perform more poorly on the retest occurring 15-min after the first administration. Although we caution that it was not significant at the level of $p < .05$, it is important to consider the effects of using multiple measures in a single test session. Among adults, a large literature has shown that participants often perform more poorly on a task requiring self-control or persistence if it was preceded by another self-control task (e.g., Baumeister, Bratslavsky, Muraven, & Tice, 1998). This effect, however, has not been reported among children. Further research investigating fatigue effects from multiple tests of EF is needed with children.

Additional research also is needed to assess longer delays between test and retest on these and other EF measures. Our conclusions are clearly limited to same-day retest, however, we predict similar results with a one-week delay, with the exception that any fatigue effects would disappear following rest. Even longer delays would need to be viewed cautiously as a measure of test-retest reliability because of the nature of rapid developmental change (at least on the Conflict tasks) in the preschool period. Nevertheless, the present research showing adequate reliability of EF tasks lends confidence to future research in this area.

Recommendations for Best Practices

This study provides an empirical foundation for recommendations for future practice. First, results suggest that the Conflict scales might be more sensitive to assessment of developmental differences compared to the Delay tasks. Therefore, the Conflict scales could be best employed when the researcher is interested in changes in EF over time. The Delay tasks, on the other hand, showed little change across age but much more variability within each age; therefore they might best be used as measures of individual differences in impulse control. However, with these recommendations one must also consider the test-retest reliability demonstrated in this paper. For example, although both Conflict-Hot and Conflict-Cool showed robust reliability, investigation of some of the subtasks showed poorer reliability, suggesting it is best to use scale scores in analyses of EF. In addition, Conflict-Cool showed a ceiling effect with 4.5-year-old children scoring below the maximum score but failing to switch dimensions flexibly on the final and most advanced subtask. Given this finding, we suggest that the DCCS Advanced-Cool might be a more sensitive assessment of EF for older children (5 to 6 years). It is unclear why Conflict-Hot did not show similar limitations.

Last, while data suggest the Delay tasks could be a more sensitive assessment of individual than developmental differences, more robust test-retest reliability was reported for Delay-Hot than Delay-Cool. We can only speculate why this might be the case. As we noted earlier, without the motivation to receive rewards themselves, over time children may alter

their performance, hence contributing to the lower test-retest reliability of the task. Our recommendation is to use the Delay-Hot task as a purer measure of Delay EF when the construct of interest is control of one's own impulses, and especially individual differences therein.

Conclusion

One obstacle in the field of cognitive development has been the issue of measurement. Researchers have used a number of tasks to assess EF, many of which are not child-friendly and are sensitive to only a limited age range. Though the multi-method approach to studying EF is beneficial, very little research has been directed toward evaluating the psychometric properties of EF measures. Hence, the primary aim of this paper was to report on the test-retest reliability of assessments of EF based on measures that are widely used. Results showed that all but one EF measure (Delay-Cool) was highly reliable. Further research is warranted to assess additional psychometric properties of the EF measures employed here such as their construct and criterion validity.

Acknowledgments

In addition to the participating families, we thank Liliana Lengua for design consultation, and research assistants Petrina Lin, Elizabeth Newmark, Joy Kawamura, and Katrina Amon. This research was supported by NICHD (5R01HD051495) awarded to SMC.

References

- Aman CJ, Roberts RJ, Pennington BF. A neuropsychological examination of the underlying deficit in attention deficit hyperactivity disorder: Frontal lobe versus right parietal lobe theories. *Developmental Psychology*. 1998; 34:956–969. [PubMed: 9779742]
- Archibald SJ, Kerns KA. Identification and description of new tests of executive functioning in children. *Child Neuropsychology*. 1999; 5:115–129.
- Baumeister RF, Bratslavsky E, Muraven M, Tice DM. Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*. 1998; 74:1252–1265. [PubMed: 9599441]
- Bechara, A.; Damasio, AR.; Reuven, B. The anatomy of emotional intelligence and implications for educating people to be emotionally intelligent. In: Reuven, B.; Maree, J.; Elias, M., editors. *Educating people to be emotionally intelligent*. Westport, CT: Praeger/Greenwood; 2007. p. 273-290.
- Beck, DM.; Carlson, SM. Correspondence between executive function and parent report of child temperament; Poster presented at the annual meeting of the Cognitive Development Society; San Diego, CA. 2005 October.
- Berlin HA, Rolls ET, Kischka U. Impulsivity, time perception, emotion and reinforcement sensitivity in patients with orbitofrontal cortex lesions. *Brain*. 2004; 127:1108–1126. [PubMed: 14985269]
- Bernier A, Carlson SM, Bordeleau S, Carrier J. Relations between physiological and cognitive regulatory systems: Infant sleep regulation and subsequent executive functioning. *Child Development*. in press.
- Bernier A, Carlson SM, Whipple N. From external regulation to self-regulation: Early parenting precursors of young children's executive functioning. *Child Development*. 2010; 81:326–339. [PubMed: 20331670]
- Bialystok E, Martin MM. Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science*. 2004; 7:325–339. [PubMed: 15595373]
- Bishop DVM, Aamondt-Leeper G, Creswell C, McGurk R, Skuse DH. Individual differences in cognitive planning on the Tower of Hanoi task: Neuropsychological maturity or measurement error? *Journal of Child Psychology and Psychiatry*. 2001; 42:551–556. [PubMed: 11383971]
- Blair C, Razza RP. Relating effortful control, executive function, and false-belief understanding to emerging math and literacy ability in kindergarten. *Child Development*. 2007; 78:647–663. [PubMed: 17381795]

- Brock LL, Rimm-Kaufman SE, Nathanson L, Grimm KJ. The contributions of 'hot' and 'cool' executive function to children's academic achievement, learning-related behaviors, and engagement in kindergarten. *Early Childhood Research Quarterly*. 2009; 24:337–349.
- Carlson SM. Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*. 2005; 28:595–616. [PubMed: 16144429]
- Carlson SM, Davis AC, Leach JG. Less is More: Executive function and symbolic representation in preschool children. *Psychological Science*. 2005; 16:609–616. [PubMed: 16102063]
- Carlson SM, Mandell DJ, Williams L. Executive function and theory of mind: Stability and prediction from age 2 to 3. *Developmental Psychology*. 2004; 40:1105–1122. [PubMed: 15535760]
- Carlson SM, Meltzoff AN. Bilingual experience and executive functioning in young children. *Developmental Science*. 2008; 11:279–295.
- Carlson SM, Moses LJ. Individual differences in inhibitory control and children's theory of mind. *Child Development*. 2001; 72:1032–1053. [PubMed: 11480933]
- Carlson SM, Moses LJ, Breton C. How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*. 2002; 11:73–92.
- Carlson SM, Wang T. Inhibitory control and emotion regulation in preschool children. *Cognitive Development*. 2007; 22:489–510.
- Casey BJ, Tottenham N, Fossella J. Clinical, imaging, lesion, and genetic approaches toward a model of cognitive control. *Developmental Psychobiology*. 2002; 40:237–254. [PubMed: 11891636]
- Coolican, H. *Research Methods and Statistics in Psychology*. London: Hodder & Stoughton; 1994.
- Davis-Unger A, Carlson SM. Children's teaching: Relations to theory of mind and executive function. *Mind, Brain, and Education*. 2008; 2:128–135.
- Diamond A, Carlson SM, Beck DM. Preschool children's performance in task switching on the Dimensional Change Card Sort task: Separating dimensions aids the ability to switch. *Developmental Neuropsychology*. 2005; 28:689–729. [PubMed: 16144433]
- Eigsti I, Zayas V, Mischel W, Shoda Y, Ayduk O, Dadlani MB, et al. Predicting cognitive control from preschool to late adolescence and young adulthood. *Psychological Science*. 2006; 17:478–484. [PubMed: 16771797]
- Fan J, McCandliss BD, Sommer T, Raz A, Posner MI. Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*. 2002; 14:340–347. [PubMed: 11970796]
- Frye D, Zelazo PD, Palfai T. Theory of mind and rule-based reasoning. *Cognitive Development*. 1995; 10:483–527.
- Garon N, Bryson SE, Smith IM. Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*. 2008; 134:31–60. [PubMed: 18193994]
- Gnys JA, Willis WG. Validation of executive function tasks with young children. *Developmental Neuropsychology*. 1991; 7:487–501.
- Grant DA, Berg EA. A behavioral analysis of degree of reinforcement and ease of shifting to a new response in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*. 1948; 38:404–411. [PubMed: 18874598]
- Hongwanishkul D, Happaney KR, Lee W, Zelazo PD. Assessment of hot and cool executive function in young children: Age-related changes and individual differences. *Developmental Neuropsychology*. 2005; 28:617–644. [PubMed: 16144430]
- Hughes C. Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology*. 1998; 16:233–253.
- Hughes C, Ensor R. Theory of mind and executive function: A family affair? *Developmental Neuropsychology*. 2005; 28:645–668. [PubMed: 16144431]
- Ingram F, Greve KW, Ingram PTF, Soukup VM. Temporal stability of the Wisconsin Card Sorting Test in an untreated patient sample. *British Journal of Clinical Psychology*. 1999; 38:209–211. [PubMed: 10389602]
- Kochanska G, Murray KT, Harlan ET. Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology*. 2000; 36:220–232. [PubMed: 10749079]

- Lowe C, Rabbitt P. Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. *Neuropsychologia*. 1998; 36:915–923. [PubMed: 9740364]
- McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. 1996; 1:30–46.
- Metcalfe J, Mischel W. A hot/cool system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*. 1999; 106:3–19. [PubMed: 10197361]
- Moriguchi Y, Hiraki K. Neural origin of cognitive shifting in young children. *Proceedings of the National Academy of Sciences*. 2009; 106:6017–6021.
- Overman WH. Sex differences in early childhood, adolescence and adulthood on cognitive tasks that rely on orbital prefrontal cortex. *Brain and Cognition*. 2004; 55:134–147. [PubMed: 15134848]
- Portney, LG.; Watkins, MP. *Foundations of clinical research: Applications to practice*. Norwalk, CT: Appleton & Lange; 1993.
- Prencipe A, Zelazo PD. Development of affective decision-making for self and other. Evidence for the integration of first- and third-person perspectives. *Psychological Science*. 2005; 16:501–505. [PubMed: 16008779]
- Qu L, Zelazo PD. The facilitative effect of positive stimuli on 3-year-old's flexible rule use. *Cognitive Development*. 2007; 22:456–473.
- Rodriguez ML, Mischel W, Shoda Y. Cognitive person variables in the delay of gratification of older children at risk. *Journal of Personality and Social Psychology*. 1989; 57:358–367. [PubMed: 2760808]
- Roid, GH. *Stanford-Binet Intelligence Scales for Early Childhood, Fifth Edition, Manual*. Itasca, IL: Riverside Publishing; 2005.
- Shoda Y, Mischel W, Peake PK. Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology*. 1990; 26:978–986.
- Tate RL, Perdices M, Maggioletto S. Stability of the Wisconsin Card sorting Test and the determination of the reliability of change in scores. *The Clinical Neuropsychologist*. 1998; 12:348–357.
- Thompson C, Barresi J, Moore C. The development of future-oriented prudence and altruism in preschoolers. *Cognitive Development*. 1997; 12:199–212.
- Zelazo PD. The dimensional change card sort (DCCS): A method of assessing executive function in children. *Nature Protocols*. 2006; 1:297–301.
- Zelazo, PD.; Carlson, SM.; Kesek, A. Development of executive function in childhood. In: Nelson, C.; Luciana, M., editors. *Handbook of Developmental Cognitive Neuroscience*. Cambridge, MA: MIT Press; 2008. p. 553-574.
- Zelazo, PD.; Cunningham, WA. Executive function: Mechanisms underlying emotion regulation. In: Gross, James J., editor. *Handbook of Emotion Regulation*. New York: Guilford Press; 2007. p. 135-158.
- Zelazo PD, Müller U, Frye D, Marcovitch S. The development of executive function in early childhood. *Monographs of the Society for Research in Child Development*. 2003; 68(3) Serial No. 274.
- Zelazo, PD.; Qu, L.; Kesek, A. Hot executive function: Emotion and the development of cognitive control. In: Calkins, S.; Bell, MA., editors. *Child development at the intersection of emotion and cognition*. Washington, DC: American Psychological Association; 2009. p. 97-111.

Figure 1.
Administration Protocol for the Conflict Scale

Figure 2.
Scatterplots of scores on test and retest for each task
Note. $** p < .01$, $*** p < .001$.

Table 1

Means, standard deviations, and minimum-maximum scores on test and retest as a function of task and age

	Test				Retest			
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>Min-Max</i>	<i>M</i>	<i>SD</i>	<i>Min-Max</i>	
Conflict-Hot								
All ages	38	24.95	15.55	4-47	24.29	15.41	3-48	
2.5 year-olds	12	9.92	10.42	4-41	9.17	6.99	3-28	
3.5 year-olds	13	23.77	11.67	12-42	22.23	11.17	12-46	
4.5 year-olds	13	40.00	6.07	23-47	40.31	7.19	21-48	
Conflict-Cool								
All ages	40	24.78	15.69	2-46	23.17	15.69	1-43	
2.5 year-olds	14	11.07	7.79	2-28	10.14	7.78	1-26	
3.5 year-olds	12	20.75	12.95	6-44	16.75	10.06	7-42	
4.5 year-olds	14	41.93	1.44	40-46	41.71	0.82	40-43	
Delay-Hot								
All ages	35	4.14	2.93	0-9	3.89	3.09	0-9	
2.5 year-olds	12	3.75	1.54	1-6	2.83	1.90	0-5	
3.5 year-olds	11	5.00	2.90	0-9	4.55	2.88	0-9	
4.5 year-olds	12	3.75	3.93	0-9	4.33	4.07	0-9	
Delay-Cool								
All ages	38	4.34	3.00	0-9	4.47	3.16	0-9	
2.5 year-olds	13	4.38	2.84	0-9	4.00	2.45	1-9	
3.5 year-olds	12	3.50	2.91	0-9	3.67	3.26	0-9	
4.5 year-olds	13	5.08	3.25	0-9	5.69	3.54	0-9	

Note: Total possible points on Conflict = 48, Delay = 9.

Table 2

Intraclass correlations for test-retest performance on each subtask of the Conflict scales.

	Hot (<i>n</i> = 38)	Cool (<i>n</i> = 40)
Categorization/Reverse Categorization	.69***	.82***
DCCS Separated	.89***	.78***
DCCS Integrated	.79***	.94***
DCCS Advanced	.58***	.90***

 $p < .001$