Published in final edited form as:
Psychol Assess. 2012 March ; 24(1): 226–239. doi:10.1037/a0025361.

The Measurement of Executive Function at Age 5: Psychometric Properties and Relationship to Academic Achievement

Michael T. Willoughby[1], **Clancy B. Blair**[2], **R.J. Wirth**[3], **Mark Greenberg**[4], and **the Family Life Project Investigators**

[1]UNC-Chapel Hill

[2]New York University

[3]Vector Psychometric Group

[4]Prevention Research Center

Abstract

This study examined the psychometric properties and criterion validity of a newly developed battery of executive function tasks for use in early childhood. The battery was included in the Family Life Project (FLP), a prospective longitudinal study of families who were over-sampled from low income and African American families at the birth of a new child (N = 1292). Ninety-nine percent (N = 1036) of children who participated in the age 5 home visit completed one or more (*M* = 5.8, *Median* = 6) of the six executive function tasks. Results indicated that tasks worked equally well for children residing in low and not low income homes, that task scores were most informative about the ability level of children in the low-average range, that performance on executive function tasks was best characterized by a single factor, and that individual differences on the EF battery were strongly related to a latent variable measuring overall academic achievement, as well as with individual standardized tests that measured phonological awareness, letter-word identification, and early math skills.

Keywords

Executive Function; Academic Achievement; Early Childhood; Psychometrics

Executive function (EF) is an umbrella term that refers to wide range of cognitive abilities—including attentional control, cognitive flexibility, goal setting, and specific aspects of information processing including fluency and processing speed—that are involved in the control and coordination of information in the service of goal-directed actions (Anderson, 2002; Fuster, 1997; Miller & Cohen, 2001). As such, EF can be defined as a supervisory system that is important for planning, reasoning ability, and the integration of thought and action (Shallice & Burgess, 1996). At a more fine grained level, however, EF, as studied in

Correspondence should be sent to Michael Willoughby, FPG Child Development Institute, UNC-CH, Campus Box 8185, 521 South Greensboro Street, Carrboro, NC 27510. willoughby@unc.edu.
Michael T. Willoughby, FPG Child Development Institute, University of North Carolina at Chapel Hill; Clancy B. Blair, Department of Applied Psychology, New York University; R.J. Wirth, Vector Psychometric Group; Mark Greenberg, Prevention Research Center; and the Family Life Project Investigators.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/PAS the cognitive development literature, has come to refer to specific interrelated information processing abilities that enable the resolution of conflicting information; namely, *working memory*, defined as the holding in mind and updating of information while performing some operation on it; *inhibitory control*, defined as the inhibition of prepotent or automatized responding when engaged in task completion; and *attention shifting*, defined as the ability to shift cognitive set among distinct but related dimensions or aspects of a given task (Davidson *et al.*, 2006; Miyake *et al.*, 2000; Zelazo & Müller, 2002). Focusing on these more narrowly defined abilities is particularly apropos when studying EF in early childhood, as many of the more complex aspects of EF (e.g., abstract thought; goal setting) have an extended developmental course and are not easily measured in very young children (Garon *et al.*, 2008).

One of the reasons for widespread interest in EF is its relation to brain development. Executive functions are closely associated with prefrontal cortex (Miller & Cohen, 2001; Stuss & Knight, 2002), an area of the brain with a protracted developmental timetable extending into early adulthood (Toga *et al.*, 2006). In keeping with knowledge of prefrontal cortex (PFC) development, cross-sectional studies have shown that performance on relatively low demand EF tasks asymptotes early, with mature levels of performance reached at approximately mid-childhood (Luciana & Nelson, 1998). Performance on tasks with more complex processing demands, however, continues to increase throughout adolescence, reaching mature levels only in young adulthood (Davidson et al., 2006; Luciana *et al.*, 2005). The relatively slow time course of PFC development and close relation between EF task performance and age has focused interest on the timing and nature of early experiences that might influence PFC development and executive functioning. As such, an important goal of research on EF has been to identify tasks that are appropriate for young children that can be used to determine expected levels of performance in order to gauge the effect of specific conditions or experiences on the development of EF and self-regulation (Blair *et al.*, 2005).

Along with Zelazo's pioneering research on the measurement of EF in young children (Zelazo & Reznick, 1991), Espy and colleagues were among the earliest to develop EF measures intended explicitly for use with young children (Espy, 1997; Espy & Cwik, 2004; Espy *et al.*, 1999a; Espy *et al.*, 2001; Espy *et al.*, 1999b). Their early work was (and remains) noteworthy in at least three ways. First, it occurred during a historical transition point characterized by a changing views about whether EF abilities were present, let alone measurable, in young children (Becker *et al.*, 1987; Fletcher & Taylor, 1984; Passler *et al.*, 1985). Second, some of their tasks were adapted from the animal neuroscience literature in which brain-behavior relationships were readily established (Espy et al., 2001); others represented careful downward extensions of adult neuropsychological tasks for use with young children (Espy & Cwik, 2004). Common to all of their task development efforts was an emphasis on developmentally appealing and appropriate test stimuli and the use of a task battery composed of multiple short tasks that accommodated the limited attention span of young children. Third, participants were drawn from birth announcements or defined populations (e.g., prenatal exposure to cocaine). These sampling strategies resulting in better external validity (in case of birth record recruitment) and more focused understanding of target populations (in the case of prenatal exposure to cocaine) than was possible from studies that relied on convenience samples.

In the 10–15 years since these early efforts at measure development, the number of new tasks designed to measure EF in early childhood has proliferated. Carlson (2005) provided a synopsis and empirical comparison of 24 such tasks. She made three observations that were relevant to our measurement development efforts. First, numerous tasks exhibited "binary distributions" and were characterized as primarily informing "pass/fail" distinctions in ability. Although pass/fail distinctions are appropriate for some questions, they mask individual differences in EF ability, which are frequently of interest. Moreover, categorizing child performance undermines statistical tests of association between executive function tasks with predictor and/or outcome variables (MacCallum *et al.*, 2002; Maxwell & Delaney, 1993). Binary and bimodal distributions may result from typical EF tasks having too few items, items that do not sufficiently vary in difficulty, and/or rapidly changing ability levels among young children. New task development efforts should attend to these issues. Second, tasks differ in terms of their difficulty level. Hence, interest in between and/or within (longitudinal change) group comparisons may be conditional on the specific tasks that were chosen. Efforts to develop new tasks should represent a broad range of ability level in order to be maximally useful for a wide variety of potential uses. Third, many tasks were developed for use in laboratory settings. As such, they have often been administered by highly trained staff (e.g., graduate students) and may involve uncommon test materials that are not easily reproducible. This may limit their wide-scale use by lay interviewers in the context of large-scale studies. Developing tasks that provide highly scripted instructions, which are amenable to use with lay data collectors, and that present tasks in a uniform and portable format may fill a practical research need.

We have been working to develop a new battery of EF tasks that attend to these limitations. Specifically, in line with the pioneering work of Espy, Diamond, and Zelazo among others, we have adapted six tasks that putatively measured three dimensions of EF—working memory, inhibitory control, and attention shifting—that may or may not be distinguishable in early childhood. The primary goal of this study was to evaluate the psychometric properties of six EF tasks that were administered to an epidemiologically-derived sample of children when they were approximately 5-years old. We previously reported psychometric properties for a subset of these tasks, in this *Journal*, when this sample was approximately 3-years old (Willoughby *et al.*, 2010). Briefly, we reported (1) that 91% of all 3-year-olds were able to complete one or more of the EF tasks, (2) that the battery of EF tasks was best represented by a single latent factor (unitary structure), and (3) that the tasks did a better job measuring lower (less proficient) than higher (more proficient) levels of executive function ability. Given children's older age in the current study, we hypothesized that 100% (vs. 91%) of children would be able to complete one or more tasks. Moreover, despite the fact that research with elementary-aged children through adult samples has indicated that EF is a multi-dimensional construct (e.g., Miyake *et al.*, 2000), EF abilities appear to be undifferentiated across the early childhood period (Espy *et al.*, 2010; Hughes *et al.*, 2010; Shing *et al.*, 2010; Wiebe *et al.*, 2008). Despite the heterogeneity of the skills required by the different tasks, we hypothesized that the EF task battery would continued to be best fit by a single factor model. This is consistent with previously reported analyses which indicated that a 1-factor model provided an optimal fit (i.e., a 2-factor model did provide a statistically significant improvement in fit relative to the 1-factor model) to a subset of these tasks that were administered at an earlier assessment (Willoughby et al., 2010). Finally, changes to the assessment battery between the 3- and 5-year old assessments included the addition of items to some tasks (i.e., the working memory span task was increased from 11 to 19 items), the complete modification of others (i.e., the spatial conflict arrows task replaced the previous spatial conflict task), and the inclusion of one entirely new task (i.e., pick the picture task). An open question was whether these changes resulted in any improvements with respect to the precision of measurement of EF abilities in the above average to superior range of latent (true) ability level.

A secondary goal of this study was to evaluate the criterion validity of the EF task battery by relating it to children's performance on standardized tests of academic achievement. EF has been implicated as an important predictor of school readiness (Blair, 2002). Individual differences in EF abilities in early childhood are associated with increased levels of prosocial and decreased levels of disruptive behavior, as well as enhanced academic achievement (Bierman *et al.*, 2009; Brock *et al.*, 2009; Smith-Donald *et al.*, 2007; Thorell & Wahlstedt, 2006; Welsh *et al.*, 2010). EF is more strongly related to academic than behavioral functioning. Moreover, EF appears to be more strongly related to math than reading achievement, which is interesting given the presumed involvement of the prefrontal cortex in both solving math problems and completing inhibitory control tasks (Blair & Razza, 2007; Bull *et al.*, 2008; Bull & Scerif, 2001; Espy *et al.*, 2004). A number of studies that have used the Head-Toes-Knees-Shoulders task, a direct assessment of child self-regulation that is conceptually similar to EF, have also demonstrated associations with academic achievement and learning related social skills (Matthews *et al.*, 2009; McClelland *et al.*, 2007; Ponitz *et al.*, 2009).

Collectively, these results suggest that efforts to directly improve EF in early childhood may facilitate pre-academic achievement, which may be particularly important for low-income and other at-risk children (Bierman *et al.*, 2008; Blair & Diamond, 2008; Diamond *et al.*, 2007). Although compelling, it is noteworthy that most of the cross-sectional associations between EF and academic achievement have been of modest magnitude, *r*s = .3 to .45 (Blair & Razza, 2007; Bull et al., 2008; McClelland et al., 2007). This raises questions about how much improvement in academic achievement can be expected through improvements in EF. However, the magnitude of these associations is ambiguous. The constructs of EF and academic achievement may only be modestly related. Alternatively, these constructs may be strongly related but the reported associations may be attenuated by poor measurement, especially of EF tasks. Indeed, four studies that used factor analysis to create EF composite scores reported stronger associations between EF and academic achievement than did comparable studies that relied on scores from individual EF tasks, *r*s = .4 – .6 (Brock et al., 2009; Espy et al., 2004; Welsh et al., 2010; Willoughby *et al.*, 2011). The current study tested the strength of the association between EF and academic achievement using a latent variable approach, which permitted an error-free estimate of the association between these constructs.

In sum, the primary objectives of this study were to evaluate the psychometric properties of each of the EF tasks that were administered at the age 5 year assessment, to test the factor structure of the overall task battery, and to test the criterion validity of the task battery with respect to children’s performance on norm-referenced tests of academic achievement. We hypothesized that the task battery would be tolerable for all children, that tasks would elicit reliable individual differences in children’s EF abilities, that children’s performance on all tasks would be best represented by a single latent factor, and that the magnitude of the association between EF and academic achievement would be larger than any of the correlations reported to date (i.e., greater than .60), due to our focus on latent correlations.

Method

Participants

The Family Life Project (FLP) was designed to study young children and their families who lived in two of the four major geographical areas of the United States with high poverty rates (Dill, 2001). Specifically, three counties in Eastern North Carolina and three counties in Central Pennsylvania were selected to be indicative of the Black South and Appalachia, respectively. The FLP adopted a developmental epidemiological design in which sampling procedures were employed to recruit a representative sample of 1292 children whose families resided in one of the six counties at the time of the child’s birth. Low-income families in both states and African American families in NC were over-sampled (African American families were not over-sampled in PA because the target communities were at least 95% non-African American).

At both sites, recruitment occurred seven days per week over the 12-month recruitment period spanning September 15, 2003 through September 14, 2004 using a standardized script and screening protocol. The coverage rate was over 90% for all births that occurred to women in these counties in that one year period. In PA, families were recruited in person from three hospitals. These three hospitals represented a weighted probability sample (hospitals were sampled proportional to size within county) of seven total hospitals that delivered babies in the three target PA counties. PA hospitals were sampled because the number of babies born in all seven target hospitals far exceeded the number needed for purposes of the design. In NC, families were recruited in person and by phone. In-person recruitment occurred in all three of the hospitals that delivered babies in the target counties. Phone recruitment occurred for families who resided in target counties but delivered in non-target county hospitals. These families were located through systematic searches of the birth records located in the county courthouses of nearby counties.

FLP recruiters identified 5471 (59% NC, 41% PA) women who gave birth to a child in the 12-month period. A total of 1515 (28%) of all identified families were determined to be ineligible for participation for three primary reasons: not speaking English as the primary language in the home, residence in a non-target county, and intent to move within three years. Of the 2691 eligible families who agreed to the randomization process, 1571 (58%) families were selected to participate using the sampling fractions that were continually updated from our data center. Of those families selected to participate in the study, 1292 (82%) families completed a home visit at 2 months of child age, at which point they were formally enrolled in the study. Interested readers are referred to a monograph that is currently submitted for publication (available upon request) that summarizes study recruitment strategies and provides detailed descriptions of participating families and their communities (Vernon-Feagans, Cox, and the Family Life Project Key Investigators, 2011).

The current study focused on children’s performance on a newly developed battery of Executive Function tasks that were administered at the age 5 year home visit, as well as their performance on multiple standardized tests of academic achievement that were administered in a prekindergarten (PreK) visit. Families and children who participated in the age 5 home visit (N = 1091) did not differ from those who did not participate in this visit (N = 201) with respect to child race (43% vs.40 % African American, $\chi^2_{(1)} = 0.5$, $p = .49$), child gender (51% vs. 53% male, $\chi^2_{(1)} = 0.3$, $p = .56$), state of residence (40% vs. 39% residing in PA, respectively, $\chi^2_{(1)} = 0.2$, $p = .67$), or being recruited in the low income stratum (78% vs. 76% poor, $\chi^2_{(1)} = 0.5$, $p = .47$). Similarly, children who participated in the PreK visit (N = 1009) did not differ from those who did not participate in this visit (N = 283) with respect to child race (43% vs. 40 % African American, $\chi^2_{(1)} = 1.3$, $p = .26$), child gender (50% vs. 53% male, $\chi^2_{(1)} = 0.7$, $p = .41$), or residing in a household that was recruited in the low income stratum at study entry (78% vs. 77% poor, $\chi^2_{(1)} = 0.1$, $p = .81$). However, children who participated in the PreK visit were more likely to reside in PA than children who did not participate in the PreK visit (42% vs. 33%, $\chi^2_{(1)} = 7.3$, $p = .007$).

In total, 77% (N = 992) of all study participants participated in both age 5 home and PreK visits, 8% (N = 99) participated in the age 5 home but not PreK visits, 1% (N = 17) participated in the PreK but not age 5 home visit, and 14% (N = 184) participated in neither the age 5 home nor the PreK visit. Children were, on average, 61 (*SD* = 3.1) months old at the age 5 home visit and 60 (*SD* = 3.4) months at the PreK visit. However, these descriptive statistics are misleading. No effort was made to temporally link the age 5 home and PreK visits, and many visits occurred over a wide window of time (difference in child age (in months) between visits: *M* = 0.4, *SD* = 4.4; range = 29 months; inter-quartile range = −3.5 to 2.9 months, where negative numbers refer to PreK visits that preceded the age 5 home visit).

Procedures

Families participated in one home visit when children were approximately 5 years old. During this visit, children completed the executive function tasks, followed by emotion recognition, parent-child interaction, and emotional challenge tasks. Caregivers also completed questionnaires and interviews. The visit took approximately two hours to complete. All visits took place at times that were convenient for family schedules, with most visits occurring in mid-morning or early afternoons (however some visits occurred in early evenings, per family requests).

Prior to beginning testing, research assistants (RAs) identified a quiet area in the residence to work (typically at kitchen or coffee table). Although efforts were made to reduce distractions (e.g., turn off TVs), in some households this was unavoidable. Parents were usually completing questionnaires at the time of child testing and did not assist with child participation with EF tasks. Efforts were made to ensure that no other persons (including siblings) in the household assisted or otherwise distracted children during testing. One RA was responsible for administering EF tasks (in a fixed order) to children, including keeping them engaged and making decisions about how frequently to take breaks. A second RA was responsible for recording children's responses to each task into a laptop computer. At the conclusion of each task, this second RA also rated the quality of the testing, including impressions of the child's comprehension of the task and the conduciveness of the home for testing. By disassembling administration and response recording roles, and not requiring either RA to evaluate the accuracy of child responses (accuracy was evaluated using computerized scoring), we minimized the demand on RAs, making the tasks more amenable to administration by lay staff who did not have specialized training or expertise in task content. Two strategies were used to ensure coordination of administration and scoring procedures. First, each page of the flip book had a clearly labeled item number that the RA who was recording responses could match against the data entry field in their laptop. Second, the RA who recorded responses would (discretely) call out the item number a few times for each task to allow the person who was administering the task to cross check the item that was being recorded. Computer scored, item-level data formed the basis of psychometric analyses included herein. EF assessments typically took 25–45 minutes (4–7 minutes per task) to complete. Variation in task length depended on how many practice trials were necessary for a child to understand the nature of the task, as well as the number of breaks that were administered (at the discretion of the RA working with the child).

The PreK visit was intended to assess children's academic achievement prior to enrollment in Kindergarten. Of the 1009 children who completed a PreK visit, 82% (N = 826) were enrolled center based care, while 4% (N = 38) received non-parental, home-based care and 14% (N = 145) received parental, home based care. PreK visits were completed in centers or homes contingent on the child's care arrangement.

Measures

Each of the six executive function tasks was presented in an open spiral bound flipbook with pages that measured 8″ × 14″. For each task, RAs first administered training trials and up to three practice trials if needed. If children failed to demonstrate an understanding of the goals of the task following the practice trials, the examiner discontinued that task.

Working Memory Span (WM)—This task is based upon principles described by Engle, Kane and collaborators (e.g., Conway *et al.*, 2005). In this task, children are presented with a line drawing of an animal figure above which is colored dot. Both the animal and the colored dot are located within the outline of a house. After establishing that the child knows both colors and animals in a pretest phase, the examiner asks the child to name the animal and then to name the color. The examiner then turns the page which only shows the outline of the house from the previous page. The examiner then asks the child which animal was/ lived in the house. The task requires children to perform the operation of naming and holding in mind two pieces of information simultaneously and to activate the animal name while overcoming interference occurring from naming the color. Children received one 1-house trial, two 2-house, two 3-house, and two 4-house trials.

Pick the Picture (PTP)—This is a Self-Ordered Pointing task (Cragg & Nation, 2007; Petrides & Milner, 1982). Children are presented with a set of pictures. For each set, they are instructed to pick each picture so that all of the pictures "get a turn". For example, in the 2-picture condition, they might see a page with pictures of an apple and a dog. On the first page, they pick (touch) either of the two pictures (child preference). On the second page, the same two pictures are presented but in a different order. Children are instructed to pick a different picture so that each picture "gets a turn". Children received two each of 2-, 3-, 4-, and 6-picture sets. The arrangement of pictures within each set is randomly changed across trials so that spatial location is not informative. This task requires working memory because children have to remember which pictures in each item set they have already touched. The person scoring the task only records which picture the child touched on each trial. Due to the dependence of responses within each picture set, each picture set is scored as a single ordinal item that reflects the number of *consecutive* correct responses beginning at the second picture of any given set (because the first picture in any set serves as a reference picture against which all responses are judged).

Spatial Conflict Arrows (SCA)—The SCA is a Simon task similar to that used by Gerardi-Caulton (2000) that is intended to assess inhibitory control. A response card, which has two side-by-side black circles that are referred to as "buttons", is placed in front of the child. The RA turns pages that depict either a left pointing or right pointing arrow. The child is instructed to touch the left most button with his/her left hand when the arrow points to the left and to touch the right most button with his/her right hand when the arrow points to the right. Across the first 8 trials, arrows are depicted centrally (in the center of the page). These items provide an opportunity to teach the child the task (touch the button left button when you see left pointing arrows and the right button when you see right pointing arrows). For items 9–22, left and right pointing arrows are depicted laterally, with left pointing arrows always appearing on the left side of the flip book page (left arrows appear "above" the left button) and right pointing arrows always appearing on the right side of the flip book page (right arrows appear "above" the right button). These items build a prepotency to touch the response card based on the location of the stimuli. For items 23–35, left and right pointing arrows begin to be depicted contra-laterally, with left pointing arrows usually (though not exclusively) appearing on the right side of the flip book page ("above" the right button of the response card) and right pointing appears appearing on the left side of the flip book page ("above" the left button of the response card). Items presented contra-laterally require inhibitory control from the previously established pre-potent response in order to be answered correctly (spatial location is no longer informative).

Something's the Same (STS)—This task, which was modeled on the Flexible Item Selection Task developed by Jacques and Zelazo (2001), is intended to assess attention shifting. In the version of the task developed for flipbook administration, children are first presented with a page on which there are two line drawn items that are similar in terms of shape, size or color. The examiner draws the child's attention to the dimension along which the items are similar, stating "See, here are two pictures. These pictures are the same, they are both (cats, blue, big, etc.)". The examiner then flips a page which presents the same two items again, to the right of which is a dashed vertical line and a picture of a third item. The new third item is similar to one of the first two items along a second dimension that is different from the similarity of the first two items. For example, if the first two items were similar in terms of shape, the third item would be similar to one of the first two items in terms of either size or color. When presenting the new, third item to the child the examiner states to the child, "See, here is a new picture. The new picture is the same as one of these two pictures. Show me which of these two pictures is the same as this new picture?" This task is preceded by a pretest in which children demonstrate knowledge of color, shape, and size.

Silly Sounds Stroop (SSS)—This task, which was modeled after the Day-Night task by Gerstadt, Hong, & Diamond (1994), is intended to assess inhibitory control of a pre-potent response. In this task children are instructed to make the sound of a dog when shown a line drawing of a cat, and to make the sound of a cat when shown a line drawing of a dog. Following a pretest phase, children are presented with 18 trials (pages) involving a line drawing of a dog and cat in random order. Due to a high degree of local dependence, only the first animal on each page is used for purposes of scoring.

Animal Go No-Go (GNG)—This is a standard go no-go task (e.g., Durston et al., 2002) that is intended to assess inhibitory motor control. Children are presented with a large button that makes a "clicking" sound when it is pressed. Children are instructed to click their button every time that they see an animal except when that animal is a pig. The examiner flips pages at a rate of one page per two seconds, with each page depicting a line drawing of 1 of 7 possible animals. The task presents varying numbers of go trials prior to each no-go trial, including, in standard order, 1-go, 3-go, 3-go, 5-go, 1-go, 1-go, and 3-go trials.

Woodcock-Johnson III Tests of Achievement (WJ III; Woodcock et al., 2001): The WJ III is a co-normed set of tests for measuring general scholastic aptitude, oral language, and academic achievement. The Letter Word Identification subtest was used as an indicator of early reading achievement, while the Applied Problems and Quantitative Concepts subtests were used as indicators of early math achievement. The validity and reliability of the WJ III tests of achievement have been established elsewhere (Woodcock et al., 2001).

Test of Preschool Early Literacy (TOPEL; Lonigan et al., 2007): The TOPEL is a norm-referenced test that was designed to identify students in Pre-Kindergarten who might be at risk for literacy problems that affect reading and writing. The Phonological Awareness subtest of the TOPEL was used in this study as an indicator of early reading achievement. The validity and reliability of the Phonological Awareness subtest has been established elsewhere (Lonigan et al., 2007)

Early Childhood Longitudinal Program Kindergarten (ECLS-K) Math Assessment (http://nces.ed.gov/ecls/kinderassessments.asp)—The ECLS-K direct math assessment was designed to measure conceptual knowledge, procedural knowledge, and problem solving within specific content strands using items drawn from commercial assessments with copyright permission, and other National Center for Educational Statistics (NCES) studies (e.g., NAEP, NELS:88). The math assessment involves a two-stage adaptive design; all children are asked a common set of "routing" items, and their performance on these items informs the difficulty level of the item set that is administered following the completion of routing items. This approach minimizes the potential for floor and ceiling effects. IRT methods were used to create math scores, using item parameters that were published in a NCES working paper that reported the psychometric properties of the ECLS-K assessments (Rock & Pollack, 2002).

Analytic Strategy

Analyses proceeded in four phases. First, confirmatory factor analyses (CFAs) were used to evaluate the dimensionality of each EF task. Each task was developed to be unidimensional. However, when the fit of unidimensional models was poor, bi-factor models were considered. Bi-factor models introduce method factors that take into account residual correlations that exist between items, even after accounting for their covariation due to a shared general factor. Second, Item Response Theory (IRT) models were applied to each task for purposes of examining differential item functioning (i.e., evaluating whether items worked equivalently for children recruited in the low and not low income strata), item parameter estimation, task scoring (i.e., computation of expected a posteriori [EAP] estimates), and for evaluating score reliabilities (a function of test information). Third, CFA models were used to evaluate the dimensionality of the entire EF battery, using EAPs from IRT models as the indicators. Fourth, CFAs were used to evaluate the criterion validity of the EF task battery by relating a latent variable representing child performance on the EF task battery to a latent variable representing child performance on academic achievement tests, as well as with each standardized test score on its own (manifest correlations). IRT models were estimated using methods outlined by Gibbons and Hedeker (1992, see also, Gibbons et. al. 2007) and as implemented in the IRTPro (Cai, du Toit, & Thissen, forthcoming) software developed as part of SBIR# HHSN-2612007-00013C. CFA models were estimated using M*plus* version 5 (Muthén & Muthén, 1998–2007). CFAs took into account the complex sampling design (stratification, over-samping of low income and, in NC, African American families).

Results

Sample Description & Rates of Executive Function Task Completion

Of the N = 1091 families who completed the age 5 home visit, 96% (N = 1045) of children had an opportunity to complete EF tasks. Children who did not have an opportunity to complete tasks either (1) moved out of the study area (defined as a 200 mile radius from participating counties), in which case family interviews were conducted by phone, or, less frequently, (2) were unexpectedly not at home during the scheduled home visit. Children who were given an opportunity to complete EF tasks were indistinguishable from children who were not given an opportunity to complete tasks with respect to child race (43% vs. 48% African American, $\chi^2_{(1)} = 0.5, p = .49$), child gender (50% male vs. 52% male, $\chi^2_{(1)} = 0.1, p = .82$), residing in a household that was recruited in the low income stratum at study entry (78% vs. 76% poor, $\chi^2_{(1)} = 0.1, p = .76$), having a primary caregiver who was married (58% vs. 63%, $\chi^2_{(1)} = 0.4, p = .53$), or having a primary caregiver with a 4-year (or higher) college degree, (17% vs. 17%, $\chi^2_{(1)} = 0.0, p = .94$). In contrast, children who were given an opportunity to complete EF tasks were more likely to reside in PA than were children who were not given an opportunity (42% vs. 9%, $\chi^2_{(1)} = 20.1, p < .0001$). This reflected a state difference in residential mobility, with families in NC being more likely to relocate out of target counties. Descriptive statistics of the families and children who participated in the age 5 year home visit, subdivided by EF completion status, are summarized in Table 1.

Of the N=1045 children who were given an opportunity to complete EF tasks, 99% (N = 1036) of children were able to complete one or more EF tasks, while 1% (N = 9) of children were unable to complete any EF task. Interviewer notes indicated that children who were unable to complete tasks typically did not understand what was being asked of them and/or had disabilities that prohibited their participation (e.g., visual impairment, cerebral palsy, autism). Among the children who completed at least one EF task, most were able to complete all six EF tasks (*M* = 5.8, *Median* = 6). Specifically, 88% completed all 6 tasks, 8% completed 5 tasks, 2% completed 4 tasks, and 2% completed between 1 and 3 tasks.

Rates of individual task completion were uniformly high: Spatial Conflict Arrows (99.5% completion), Item Selection (98.7% completion), Self Ordered Pointing, (96.7% completion), Working Memory Span (94.8% completion), Silly Sounds Stroop (95.9% completion), and Go No-Go (94.5% completion). Analyses using Item Response Theory (below) provide a formal evaluation of task difficulty.

Dimensionality of Individual EF Tasks

All tasks were developed to be unidimensional. Hence, initially, a 1-factor model was fit to each EF task. When model fit was poor modification indices and standardized residuals were evaluated (see, for example, Hill et al., 2007). This informed the fitting bi-factor models, which took into account plausible patterns of residual item correlations. To be clear, we used Hu and Bentler's (1999) recommendations (i.e., CFI > .95, RMSEA <= .05) as a guide for evaluating model fit. However, these cutoffs were developed for continuous data using maximum likelihood (ML) estimation. By contrast, the CFA models used here involved discrete item data and parameter estimates were obtained using a diagonally weighted least squares estimator. Given concern that Hu and Bentler's (1999) recommendations may be too stringent for these types of models (see, for example, Cook, Kallen, & Amtmann, 2009), we also considered model residuals, tests of local dependence, and stability of parameter estimates across models in which subsets of items were removed, in order to evaluate the adequacy of the models.

Spatial Conflict Arrows (SCA)—A unidimensional model containing 11 SCA items was found to fit the data poorly ($\chi^2_{(29)}$ = 523.4, p < .0001, CFI = .89, RMSEA = 0.13, N = 1033). Examination of the results suggested estimation and inter-item dependency problems with item 11 (Item 36 in the flipbook) was the primary cause of model misfit. As such, this item was dropped from the analyses. Examination of a unidimensional model containing only the first 10 items was found to fit the data moderately well ($\chi^2_{(26)}$ = 318.8, p < .0001, CFI = .93, RMSEA = 0.10, N = 1033).

Silly Sounds Stroop (SSS)—A unidimensional model containing 15 SSS items was found to fit the data poorly ($\chi^2_{(23)}$ = 1296.0, p < .0001, CFI = .68, RMSEA = 0.24, N = 995). However, a bi-factor model that included two orthogonal method factors (one for “Cat” and “Dog” items respectively) was found to fit the data moderately well ($\chi^2_{(39)}$ = 329.9, p < . 0001, CFI = .93, RMSEA = 0.09, N = 995).

Animal Go No-Go (GNG)—A unidimensional model containing 7 no-go items was found to fit the data well ($\chi^2_{(13)}$ = 15.8, p < .2626, CFI = 1.00, RMSEA = 0.02, N = 980). Although analyses of this task when children were 3 years-old indicated that a unidimensional model with all factor loadings constrained to be equal fit the data well, constraining all of the factor loadings to equality at this assessment resulted in significantly worse model fit relative to the model in which all factor loadings were allowed to freely vary ($\chi^2_{(5)}$ = 71.6, p < .0001, N = 980).

Something's the Same (STS)—A unidimensional model containing 16 STS items was found to fit the data poorly ($\chi^2_{(61)}$ = 586.4, p < .0001, CFI = .89, RMSEA = 0.09, N = 1024). However, a bi-factor model that included three orthogonal method factors (one for “Color,” “Size,” and “Object” items respectively) was found to fit the data well ($\chi^2_{(53)}$ = 169.2, p < . 0001, CFI = .98, RMSEA = 0.05, N = 995).

Working Memory Span (WM)—Preliminary analyses determined that item dependencies between successively administered items were substantial enough to warrant collapsing of items on each page, resulting in six ordinal scores (2, 2-item scores; 2, 3-item scores; and 2, 4-item scores). A unidimensional model was found to fit the 6 WM composite items moderately well ($\chi^2_{(8)}$ = 47.7, p < .0001, CFI = .91, RMSEA = 0.07, N = 983).

Pick The Picture (PTP)—Preliminary analyses also determined that item dependencies within sets of items were substantial enough to warrant collapsing of items within each set, resulting in 8 ordinal scores (2, 1-item scores; 2, 2-item scores; 2, 3-item scores, and 2, 5-item scores). A unidimensional model was found to fit the 8 PTP composite items well ($\chi^2_{(8)}$ = 50.2.1, p < .0002, CFI = .98, RMSEA = 0.04, N = 1002).

Item Response Theory (IRT): Differential Item Functioning & Item Parameter Estimation

Unidimensional and bi-factor IRT structures were parameterized using either the 2-parameter logistic model (2PLM) or Samejima's (1969) graded response model (GRM). The 2PLM was applied to four of the six tasks with dichotomous responses (i.e., SCA, SSS, GNG, STS). The GRM was applied to the WM and PTP tasks, because, as described above, the item dependencies between dichotomous responses were collapsed into ordinal scores. Before final item parameter estimation was conducted, each item was assessed for differential item functioning for children who were recruited into the low and not low income strata using nested log-likelihood tests. Due to the large number of tests, the false discovery rate of .05 was controlled using the Benjamini-Hochberg method (see, Benjamini & Hochberg, 1995). The difficulty and discrimination parameters for items from each task are reported below.

To facilitate interpretation of the sections below for readers who are less familiar with IRT, difficulty and discrimination parameters are conceptually similar to item intercepts and factor loadings, respectively. Difficulty can be interpreted as if on a z score metric, where negative values indicate easy items, values near 0 indicate average difficulty, and positive values indicate difficult items. Discrimination parameters can be interpreted as slopes, where higher values suggest a stronger relationship to the latent variable (in practice, using the model parameterizations used here, values greater than 1 are desirable).

Spatial Conflict Arrows (SCA)—No significant DIF was found for the SCA items after controlling the false discovery rate. On average, SCA items were relatively easy for children to complete (difficulty parameters M = −.67, SD = .47, range −1.36 – 0.00). Children's performance on SCA items were moderately to strongly related to underlying ability (discrimination parameters M = 2.0, SD = 0.8, range 0.93 – 2.96).

Silly Sounds Stroop (SSS)—No significant DIF was found for the SSS items. On average, SSS items were relatively easy for children to complete (difficulty parameters M = −1.52, SD = 1.19, range −5.27 - −0.39). There was substantial variability in the strength of the relationship between individual items and underlying ability (discrimination parameters M = 1.55, SD = 0.95, range 0.11 – 3.67).

Animal Go No-Go (GNG)—No significant DIF was found for the GNG items. The GNG items were easy for children to complete (difficulty parameters M = −1.46, SD = 0.26, range −1.88 - −1.19). The GNG items were moderately related to underlying ability (discrimination parameters M = 1.74, SD = 0.49, range 0.97 – 2.20). The variation in item discrimination parameters for the GNG was appreciably less than that observed for the other tasks.

Something's the Same (STS)—No significant DIF was found after controlling for the false discovery rate. A wide range of difficulty was observed for the STS items (M = −1.51, SD = 2.23, range −6.48 – 0.04), though the group of items was, on average, relatively easy for children to complete. The IS items were related to underlying ability (discrimination parameters *M* = 1.94, *SD* = 0.94, range .73 – 3.77).

Working Memory Span (WM)—No significant DIF was found for the WM items. Unlike the previous tasks which included a series of dichotomously scored items, the WM task involved ordinal scores. As expected, the difficulty parameters increased as a function of the number of items within each ordinal score that were answered correctly. However, the discrimination parameters were of modest magnitude and highly variable (*M* = 1.08, *SD* = .76, range 0.54 – 2.06). Further inspection revealed that four of the six items (items 3–6) were weakly associated with ability (discrimination parameters < 0.72). Hence, whereas the first two items were moderately related to ability, they were relatively easy for children to complete (e.g., the difficulty parameters for [two category] item 1 were −1.75 and −1.53). In contrast, the third through sixth items were only weakly related to ability, but represented the full range of difficulty (e.g., the difficulty parameters for [four category] item 6 were −2.06, 0.65, 2.89, 4.81).

Pick the Picture (PTP)—No significant DIF was found for the PTP items. Like the WM task, the PTP task consisted of ordinal items. The difficulty parameters increased as a function of the number of items within each ordinal score were answered correctly (e.g., difficulty parameters for item 4 were −1.94 and −0.77; for item 8 were −2.08, −0.64, 0.21, 1.26, 2.33). Unlike the WM task, the set of PTP items were similarly informative of underlying ability level. Like the WM task, the difficulty range for more complicated items spanned a wide range of ability level.

Item Response Theory (IRT): Score Estimation and Reliability

In the process of item parameter estimation described above, IRT-based scores (i.e., expected a posteriori [EAP] estimates of task performance) were computed to reflect each child’s performance on each task. Descriptive statistics for EF EAP scores are included in Table 2. One of the benefits of IRT is the ability to examine how well a given test performs over the range of the latent construct. While scale performance had traditionally been evaluated in terms of “test information,” it can be equivalently evaluated in terms of the reliability of the scale scores over the range of theta (i.e., latent ability). As shown in Figure 1, score reliabilities from the six EF tasks peak from between 2 standard deviations below up to the population mean of EF ability. That is, each of the scales provide the most reliable scores for children whose true ability ranges from relatively (but not extremely) poor (i.e., 2 standard deviations below the population mean) to average (at the population mean). Moreover, four of EF tasks (SCA, PTP, STS, SSS) provide scores with reliabilities greater than .7 for children who have true ability levels as high as one standard deviation above the population mean.

Dimensionality of the EF Task Battery

A confirmatory factor analysis (CFA) model using robust full information maximum likelihood (FIML) was used to test the dimensionality of the task battery. The use of robust FIML estimation ensured that any child who completed at least one EF task was included in this analysis, as well as accommodated non-normal distributions of EF task scores. A one-factor model fit the task scores extremely well ($\chi^2_{(9)}$ = 6.3, *p* = .71, CFI = 1.0, RMSEA = 0.00, RMSEA 95% CI = 0.00 – 0.03, N = 1036). The latent variance was significant indicating inter-individual differences on task performance. The factor loadings for all six tasks were statistically significantly (*p*s < .0001). The R^2 values for individual task (EAP) scores were discrepant with the single EF factor explaining 6–40% of variation in each task (Spatial Conflict Arrows: R^2 = .06; Working Memory Span: R^2 = .13; Something’s the Same: R^2 = .13; Animal Go No-Go: R^2 = .22; Silly Sounds Stroop: R^2 = .23; Pick the

Picture: $R^2 = .40$). These modest R^2 values for each task are consistent with very modest inter-correlations between tasks.

The excellent fit of the one-factor model suggested that testing the fit of two-factor model was likely unnecessary. Nonetheless, given the uncertainty regarding the dimensionality of EF in early childhood, a two-factor CFA model was estimated in which the three tasks that putatively measured inhibitory control (SCA, GNG, SSS) and the one task that putatively measured attention shifting (STS) loaded separately from a factor defined by the two working memory tasks (WM, PTP). This two-factor model fit the task scores extremely well ($\chi^2_{(8)} = 4.5$, $p = .81$, CFI = 1.0, RMSEA = 0.00, RMSEA 95% CI = 0.00 – 0.02, N = 1036). Both latent variances were statistically significant, and the factors were positively correlated ($\varphi = .89$, $p < .001$). The factor loadings for all tasks were statistically significant (ps < .0001). Relative to the one-factor model, trivial changes in R^2 values were observed for most tasks (Spatial Conflict Arrows: $R^2 = .07$; Working Memory Span: $R^2 = .13$; Something's the Same: $R^2 = .13$; Animal Go No-Go: $R^2 = .23$; Silly Sounds Stroop: $R^2 = .24$; Pick the Picture: $R^2 = .47$). Not surprisingly, the two-factor model did not provide a statistically significant improvement in model fit relative to the one-factor model, ($\chi^2_{(1)} = 1.8$, $p = .18$), as indicated by a comparison appropriate for model comparisons involving robust ML estimation (Satorra & Bentler, 1999).

The small number and nature of tasks did not permit a test of a 3-factor (inhibitory control, working memory, attention shifting) model. With only six tasks, the three-factor model would have been just-identified and could not have provided a test of model fit against the observed data (i.e., model fit would be perfect because the model would have estimated as many parameters as there were unique (co)variances between the tasks). In addition, given that only one task (STS) was a potential indicator of the attention shifting factor, the model would have suffered from local identification problems, in that we could not estimate both factor loading and residual variance parameters from a single observed variance for STS.

Criterion Validity of the EF Task Battery

Bivariate correlations between scores for the six EF tasks and the five academic achievement tests are summarized in Table 2. Whereas children's performance on individual EF tasks was very modestly correlated (rs = .06 – .31), their performance on achievement tests was moderate to strongly correlated (rs = .45 – .78). The correlations between individual EF tasks and achievement tests took on intermediate values (rs = .06 – .41). Most of the correlations between individual EF tasks and achievement scores were between .20 and .40, which is consistent with magnitude of effects that have been reported in other studies involving preschool-aged samples. Unweighted descriptive statistics that informed the central tendency of all measures are also provided in Table 2. All of the variables under consideration were (approximately) normally distributed.

A final set of CFAs were estimated to establish the criterion validity of the EF task battery by relating performance on EF tasks to performance on standardized tests of academic achievement. Prior to estimating CFAs that informed criterion validity, two competing CFA models were estimated to evaluate the factor structure of the achievement tests (N = 971). The first, one-factor model implied that academic achievement in early childhood was best considered as unidimensional. The second, two-factor model implied that early math ability (ECLS-K math, WJ Applied Problems, WJ Quantitative Concepts) could be differentiated from reading ability (WJ Letter-Word Identification, TOPEL Phonological Processing). The one factor model fit the achievement test scores well ($\chi^2_{(5)} = 13.0$, $p = .02$, CFI = 1.0, RMSEA = 0.04, RMSEA 95% CI = 0.01 – 0.07). The latent variance was statistically significant, as were the factor loadings for all five tests (ps < .0001). The R^2 values for individual achievement scores were moderate to large (ECLS-K Math: $R^2 = .73$; WJ Applied

Problems: R^2 = .64; WJ Quantitative Concepts: R^2 = .82; WJ Letter Word: R^2 = .56; Phonological Processing: R^2 = .37). The lower value for phonological awareness relative to other measures is indicative of the fact that phonological awareness is a correlate, but not necessarily direct indicator, of academic achievement. Although the two-factor model fit the data well, ($\chi^2_{(4)}$ = 12.3, p = .02, CFI = 1.0, RMSEA = 0.05, RMSEA 95% CI = 0.02 – 0.08), the results were not trustworthy due to a Heywood case indicating that the latent correlation between math and reading factors exceeded 1. This result was interpreted as over-fitting the data and the two-factor solution was not considered further (Chen *et al.*, 2001).

A two-factor (one factor each for EF and achievement scores) CFA model was estimated to inform criterion validity. A total of N = 1058 cases contributed to this analysis (i.e., anyone with non-missing data for any EF or achievement task). This two-factor model fit the task scores well ($\chi^2_{(43)}$ = 135.1, p < .0001, CFI =.96, RMSEA = 0.05, RMSEA 95% CI = 0.04 – 0.05). All of the factor loadings and latent variances were statistically significant (ps < .001). The latent factors were positively correlated (ϕ =.70, p < .001). A parallel model was also estimated in which the EF latent variable was correlated with each of the achievement scores individually (simultaneously). This model fit the data well too ($\chi^2_{(34)}$ = 76.8, p < .0001, CFI =.98, RMSEA = 0.03, RMSEA 95% CI = 0.02 – 0.05). The EF latent variable was significantly correlated with all five achievement tests ($r_{ECLS\text{-K Math}}$= .62, $r_{WJ\ Applied\ Problems}$ = .63, $r_{WJ\ Quantitative\ Concepts}$ = .62, $r_{WJ\ Letter\text{-}Word}$ = .39, and $r_{TOPEL\ Phonological}$ = .56, all ps < .001). A follow-up model was estimated that equated the covariance between the EF latent variable and the three WJ achievement tests (which shared a common scale). Although this model fit the data reasonably well ($\chi^2_{(36)}$ = 129.2, p < .0001, CFI =.96, RMSEA = 0.05, RMSEA 95% CI = 0.04 – 0.06), compared to the previous model in which these covariances were freely estimated, it provided a statistically significant worse fit it fit the data ($\chi^2_{(2)}$ = 48.7, p < .0001). This confirms that EF was more strongly associated with performance on math than (pre)reading achievement tasks (i.e., equating the larger sized correlations involving math to lower sized correlations involving letter-word identification resulted in a degradation of model fit).

Discussion

Given the relation of executive functions (EF) to a number of aspects of child development —including behavioral and academic indicators of school readiness—research on the measurement of EF in young children is a scientific priority. Increased precision in the measurement of early EF will facilitate an improved understanding of the developmental course of EF in early childhood, including the identification of naturally occurring experiences, as well as experimental interventions, that promote competence and resilience in children at risk for school failure (Blair *et al.*, 2005). With these goals in mind, this study reported the psychometric properties, factor structure, and criterion validity of a newly developed battery of tasks designed to measure executive functioning with in a population-based sample of five year old children who resided in predominantly low-income, non-urban communities.

Ninety-nine percent of children who were given an opportunity to complete the EF battery (i.e., those for whom an in-home visit at age 5 was conducted) completed one or more tasks, and a majority of children completed all six tasks. To the best of our knowledge, this is the first study that has presented EF tasks to children in a large sample selected at birth with no exclusions. Our results demonstrate that direct child assessments of EF in the context of large scale studies are feasible. Extensive pilot testing with low income families underscored our impression that tasks should be presented in a structured manner, with as minimal language demands as possible. The 99% completion rate compares favorably to the 91% completion rate that we reported in this sample at the age three assessment.

Although numerous tasks purported to measure EF in early childhood have been developed, most widely used tasks have not undergone formal psychometric evaluations, including the presentation of reliability data, in large samples of preschool-aged children; the NEPSY neuropsychological battery—which includes some assessment of EF in addition to assessments of language, visual-spatial processing, sensorimotor function, and memory and learning—and the aforementioned Shape School task are two notable exceptions (Espy et al., 2006; Korkman *et al.*, 1998). The IRT-based approach permits a consideration of how reliability differs as a function of ability level, which is not possible from the perspective of classic test theory. Specifically, reliability curves revealed that the current battery provides the most reliable measurement of EF for five year-old children whose latent ability is poor (two standard deviations below the mean) to average (at the population mean). A few tasks provided acceptable levels of reliability (>= .70) for children whose latent ability level was slightly above average (up to one standard deviation above the population mean). It is noteworthy that the pick the picture and item selection tasks had both the highest reliability at intermediate to high levels of EF ability (see Figure 1) and were more highly correlated with all five achievement tasks than were any of the remaining four EF tasks (see Table 2). The reliability of these tasks benefits from a broader range of item difficulties, relative to the other tasks. We are using reliability curves to guide our task modifications. Specifically, we are seeking to expand the range of item difficulty on all tasks in order to improve our precision of measurement of EF across a wider ability range. The routine presentation of reliability curves for measures of EF in early childhood would facilitate the selection of tasks that closely match the characteristics of children under study, as well as to enhance the interpretation of results from studies relating EF to indicators of academic and behavioral functioning. EF tasks that have been observed to exhibit markedly skewed score distributions (primarily informing ‘pass/fail’ decisions) likely exhibit reliability curves that are extremely “peaked” in a narrow ability range.

There was no evidence that any of the tasks exhibited differential item functioning for children residing in low and not-low income homes at the time of study entry (i.e., item parameter estimates could be equated without a degradation of model fit). This helps ensure that any resulting group differences in mean level performance is not due to test bias. Nonetheless, we extensively pilot tested early versions of these tasks with low income *and* young (3 year-old) children. That pilot testing revealed that tasks had to be simplified both in language and in structure in order to work with a sufficiently large number of young, low income children. Our decision to construct tasks in a way that was most amenable to assessing abilities among young, low income children may have indirectly affected task precision, as the measurement of lower levels of ability (characteristic of young children residing in low income homes) was given more attention than was the measurement of higher ability.

When scores for the entire battery were considered together, a unidimensional model was found to fit the data well. This result is consistent with previous studies that examined the dimensionality of EF in young children (Espy et al., 2010; Shing et al., 2010; Wiebe et al., 2008), as well as results from this sample at the 3 year assessment (Willoughby et al., 2010). This study adds to a growing body of work that has demonstrated that, although EF abilities may be best conceptualized as being multi-dimensional in older children and adults, they are better conceptualized as being unidimensional (undifferentiated) in early childhood. A variety of biological (e.g., brain development) and experiential (e.g., increased complexity of adult requests) factors that occur during the transition from early to middle childhood likely contribute to the shift from unidimensional to multidimensional factor structure of EF abilities. The conclusion that EF is unidimensional in early childhood is tempered by two caveats. First, studies involving older children and adults have administered a larger number of EF tasks, which provides a stronger test of dimensionality. This strategy presents unique challenges for young children (test burden) that could not be easily addressed within the context of the current study. Second, all of the studies conducted to date have implicitly assumed that the factor structure of EF tasks is equivalent for all children. It is possible that the factor structure of EF tasks may differ for distinct subgroups of children in the population (e.g., children with disabilities). Future studies might consider the application of new analytic techniques that are designed to test this type of question (Lubke & Muthen, 2005; Yung, 1997)

It is noteworthy that although the EF tasks were best represented by a single-factor model, the correlations between tasks were very modest (see Table 2). Indeed, inter-task correlations were somewhat smaller than we have observed in other studies. We believe that this was likely due to relatively narrow age range of the sample. Taken together, the modest correlations between tasks combined with the good fit of a single-factor indicate that children's performance on individual EF tasks tend to be "noisy" indicators of true (latent) ability level (i.e., performance on any given EF task contains relatively little "signal"). Confirmatory factor analysis represents a useful tool for aggregating children's performance across a variety tasks in order to obtain an estimate of true (latent) ability level.

Criterion validity analyses demonstrated that children's performance on the EF battery was strongly correlated ($\varphi = .70$) with their performance on standardized measures of academic achievement. To the best of our knowledge, this is the largest reported correlation between EF and academic achievement of any extant study involving preschool-age children. The correlation between the EF latent variable and individual achievement tasks were also moderate to large, and this was particularly true for the association of EF and math tasks. Correlations involving the *EF latent variable* and academic achievement were appreciably larger in magnitude than were the corresponding correlations between *individual EF tasks* and achievement tests (see Table 2). As we have demonstrated elsewhere, whereas individual EF task scores include a combination of measurement error and systematic variation that is idiosyncratic to a particular task, the EF battery scores (as represented by the EF latent variable) reflects systematic variation in child ability that is common across EF tasks (Willoughby & Blair, 2011). Our results suggest that most of the reported associations between EF and academic achievement in early childhood represent lower-bound estimates that were likely attenuated by low levels of reliability for individual EF tasks. Despite the strong association between EF and academic achievement reported here, it would be inappropriate to assume that this result informs questions about the role of improving EF in early childhood in an effort to facilitate academic school readiness. Questions about the causal relationship between EF and academic achievement in early childhood are best addressed using a counterfactual framework that makes use of randomized designs, natural experiments, and/or analytic strategies that attend to sample selection effects that likely contribute to the strong association between EF ability and performance on tests of academic achievement (Willoughby *et al.*, in press).

This study was characterized by at least three strengths. First, the battery was tested using a representative sample of five year old children, with over-sampling for low income in the sample as a whole and in one site, African American families. To be clear, the combined use of an explicit sampling design and corresponding analytic strategy (stratification variables, probability weights) permitted us to over-sample families of particular interest to the larger study (family poverty), while retaining the ability to generalize our results to all families residing in the 6-county sampling area, as if we had conducted a simple random sample. Second, the tasks were developed to facilitate standard administration by lay interviewers who did not have expertise in EF. High rates of task completion demonstrate the feasibility of using direct child assessments in the context of large scale studies. Third, the use of IRT methods demonstrated how the reliability of EF tasks varied as a function of child ability level. In the future, it may be beneficial to explicitly develop EF tasks that are optimized to the assessment of children with specific ability profiles.

This study is also characterized by at least four weaknesses. First, EF tasks in children's homes. Despite efforts to standardized testing procedures, many households were characterized by non-optimal test conditions (e.g., frequent interruptions from others coming in/out of the room or household, poor lighting, high levels of ambient sound), which may have differentially undermined children's task performance. Alternatively, to the extent that household structure and/or family environments and routines facilitate the development of EF abilities, home-based testing may have greater ecological validity than clinic or lab-based assessments. Second, the FLP sample is representative of two, 3-county areas that are non-metropolitan and characterized by low wealth. These counties were selected with consideration of both logistical (e.g., proximity to University research centers) and substantive (no cities larger than 50,000 people; high density of low income families) criteria. Because counties were selected, not sampled, these results are in no way nationally representative of all five year-old children. Third, this study did not test the construct validity of this battery. In the future, it will be important to test whether children's performance on this task battery corresponds to their performance on more well established EF tasks (batteries). Fourth, in its current form, the task battery requires the use of two RAs, one for administration and the other for scoring, along with specialty printed materials and computerized scoring code. This makes it impractical for more widespread use. Efforts are currently underway to computerize the task battery, which will facilitate the administration of the battery by a single RA, will standardize aspects of tasks (e.g., inter-stimulus interval) across data collection teams, and will facilitate scoring without the availability of specialized 3rd party software.

Despite an explosion of research on children's self regulation in early childhood, the field continues to be dependent on tasks that have not been subjected to rigorous psychometric evaluation. Moreover, given a central assumption that early childhood is characterized by rapid developmental onset of EF abilities, it will be imperative to develop scalable instruments that facilitate inferences about inter-individual differences in intra-individual change in EF across ages 3–5 years. This study represents our continued effort towards this goal.

Acknowledgments

Support for this research was provided by the National Institute of Child Health and Human Development grants R01 HD51502 and P01 HD39667, with co-funding from the National Institute on Drug Abuse. The Family Life Project (FLP) Phase I Key Investigators include: Lynne Vernon-Feagans, The University of North Carolina; Martha Cox, The University of North Carolina; Clancy Blair, The Pennsylvania State University; Peg Burchinal, The University of North Carolina; Linda Burton, Duke University; Keith Crnic, The Arizona State University; Ann Crouter, The Pennsylvania State University; Patricia Garrett-Peters, The University of North Carolina; Mark Greenberg, The Pennsylvania State University; Stephanie Lanza, The Pennsylvania State University; Roger Mills-Koonce, The University of North Carolina; Debra Skinner, The University of North Carolina; Emily Werner, The Pennsylvania State University and Michael Willoughby, The University of North Carolina

References

Anderson P. Assessment and development of executive function (ef) during childhood. Child Neuropsychology. 2002; 8(2):71–82. [PubMed: 12638061]

Becker MG, Isaac W, Hynd GW. Neuropsychological development of nonverbal behaviors attributed to "frontal lobe" functioning. Developmental Neuropsychology. 1987; 3(3–4):275–298.

Bierman KL, Nix RL, Greenberg MT, Blair C, Domitrovich CE. Executive functions and school readiness intervention: Impact, moderation, and mediation in the head start redi program. Development and Psychopathology. 2008; 20(3):821–843. [PubMed: 18606033]

Bierman KL, Torres MM, Domitrovich CE, Welsh JA, Gest SD. Behavioral and cognitive readiness for school: Cross-domain associations for children attending head start. Social Development. 2009; 18(2):305–323.

Blair C. School readiness - integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. American Psychologist. 2002; 57(2):111–127. [PubMed: 11899554]

Blair C, Diamond A. Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. Development and Psychopathology. 2008; 20(3): 899–911. [PubMed: 18606037]

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(2):647–663. [PubMed: 17381795]

Blair C, Zelazo PD, Greenberg MT. The measurement of executive function in early childhood. Developmental Neuropsychology. 2005; 28(2):561–571. [PubMed: 16144427]

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(3):337–349.

Bull R, Espy KA, Wiebe SA. Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. Developmental Neuropsychology. 2008; 33(3):205–228. [PubMed: 18473197]

Bull R, Scerif G. Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. Developmental Neuropsychology. 2001; 19(3):273–293. [PubMed: 11758669]

Carlson SA. Developmentally sensitive measures of executive function in preschool children. Developmental Neuropsychology. 2005; 28(2):595–616. [PubMed: 16144429]

Chen FN, Bollen KA, Paxton P, Curran PJ, Kirby JB. Improper solutions in structural equation models - causes, consequences, and strategies. Sociological Methods & Research. 2001; 29(4):468–508.

Conway ARA, Kane MJ, Bunting MF, Hambrick DZ, Wilhelm O, Engle RW. Working memory span tasks: A methodological review and user's guide. Psychonomic Bulletin & Review. 2005; 12(5): 769–786. [PubMed: 16523997]

Cragg L, Nation K. Self-ordered pointing as a test of working memory in typically developing children. Memory. 2007; 15(5):526–535. [PubMed: 17613795]

Davidson MC, Amso D, Anderson LC, Diamond A. Development of cognitive control and executive functions from 4–13 years: Evidence from manipulations of memory, inhibition, and task switching. Neuropsychologia. 2006; 44:2037–2078. [PubMed: 16580701]

Diamond A, Barnett wS, Thomas J, Munro S. Preschool program improves cognitive control. Science. 2007 November 30.318:1387–1388. [PubMed: 18048670]

Dill, BT. Rediscovering rural america. In: Blau, JR., editor. Blackwell companions to sociology. Malden: Blackwell Publishing; 2001. p. 196-210.

Espy KA. The shape school: Assessing executive function in preschool children. Developmental Neuropsychology. 1997; 13(4):495–499.

Espy KA, Bull R, Martin J, Stroup W. Measuring the development, of executive control with the shape school. Psychological Assessment. 2006; 18(4):373–381. [PubMed: 17154758]

Espy KA, Cwik MF. The development of a trial making test in young children: The trails-p. Clinical Neuropsychologist. 2004; 18(3):411–422. [PubMed: 15739812]

Espy KA, Kaufmann PM, Glisky ML. Neuropsychological function in toddlers exposed to cocaine in utero: A preliminary study. Developmental Neuropsychology. 1999a; 15(3):447–460.

Espy KA, Kaufmann PM, Glisky ML, McDiarmid MD. New procedures to assess executive functions in preschool children. The Clinical Neuropsychologist. 2001; 15(1):46–58. [PubMed: 11778578]

Espy KA, Kaufmann PM, McDiarmid MD, Glisky ML. Executive functioning in preschool children: Performance on a-not-b and other delayed response format tasks. Brain and Cognition. 1999b; 41(2):178–199. [PubMed: 10590818]

Espy KA, McDiarmid MM, Cwik MF, Stalets MM, Hamby A, Senn TE. The contribution of executive functions to emergent mathematic skills in preschool children. Developmental Neuropsychology. 2004; 26(1):465–486. [PubMed: 15276905]

Espy KA, Sheffield TD, Wiebe S, Clark CAC, Moehr M. Executive control and dimensions of problem behaviors in preschool children. Journal of Child Psychology and Psychiatry. 2010; 52(1):33–46. [PubMed: 20500238]

Fletcher JM, Taylor HG. Neuropsychological approaches to children: Towards a developmental neuropsychology. Journal of Clinical Neuropsychology. 1984; 6(1):39–56. [PubMed: 6699184]

Fuster, JM. The prefrontal cortex. Anatomy, physiology and neuropsychology of the frontal lobe. NY: Lippincott-Raven Press; 1997.

Garon N, Bryson SE, Smith IM. Executive function in preschoolers: A review using an integrative framework. Psychological Bulletin. 2008; 134(1):31–60. [PubMed: 18193994]

Hughes C, Ensor R, Wilson A, Graham A. Tracking executive function across the transition to school: A latent variable approach. Developmental Neuropsychology. 2010; 35(1):20–36. [PubMed: 20390590]

Korkman, M.; Kirk, U.; Kemp, S. Nepsy: A developmental neuropsychological assessment manual. San Antonio: Psychological Corporation; 1998.

Lonigan, CJ.; Wagner, RK.; Torgesen, JK.; Rashotte, CA. Topel: Test of preschool early literacy. Austin, TX: PRO-ED, Inc; 2007.

Lubke GH, Muthen B. Investigating population heterogeneity with factor mixture models. Psychological Methods. 2005; 10(1):21–39. [PubMed: 15810867]

Luciana M, Conklin HM, Hooper CJ, Yarger RS. The development of nonverbal working memory and executive control processes in adolescents. Child Development. 2005; 76(3):697–712. [PubMed: 15892787]

Luciana M, Nelson CA. The functional emergence of prefrontally-guided working memory systems in four- to eight-year-old children. Neuropsychologia. 1998; 36(3):273–293. [PubMed: 9622192]

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychological Methods. 2002; 7(1):19–40. [PubMed: 11928888]

Matthews JS, Ponitz CC, Morrison FJ. Early gender differences in self-regulation and academic achievement. Journal of Educational Psychology. 2009; 101(3):689–704.

Maxwell SE, Delaney HD. Bivariate median splits and spurious statistical significance. Psychological Bulletin. 1993; 113(1):181–190.

McClelland MM, Cameron CE, Connor CM, Farris CL, Jewkes AM, Morrison FJ. Links between behavioral regulation and preschoolers’ literacy, vocabulary, and math skills. Developmental Psychology. 2007; 43(4):947–959. [PubMed: 17605527]

Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience. 2001; 24:167–202.

Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. Cognitive Psychology. 2000; 41(1):49–100. [PubMed: 10945922]

Muthén, LK.; Muthén, BO. Mplus users guide. 5. Los Angeles, CA: 1998–2007.

Passler MA, Isaac W, Hynd GW. Neuropsychological development of behavior attributed to frontal lobe functioning in children. Developmental Neuropsychology. 1985; 1(4):349–370.

Petrides M, Milner B. Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. Neuropsychologia. 1982; 20(3):249–262. [PubMed: 7121793]

Ponitz CC, McClelland MM, Matthews JS, Morrison FJ. A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. Developmental Psychology. 2009; 45(3): 605–619. [PubMed: 19413419]

Rock, DA.; Pollack, JM. N. C. f. E. S. U.S. Department of Education. Early childhood longitudinal study - kindergarten class of 1998–99 (ecls-k), psychometric report for kindergarten through first grade. Vol. Working Papers Series. Washington, DC: 2002.

Shallice T, Burgess P. The domain of supervisory processes and temporal organization of behaviour. Philosophical Transactions of the Royal Society B-Biological Sciences. 1996; 351(1346):1405–1411.

Shing YL, Lindenberger U, Diamond A, Li SC, Davidson MC. Memory maintenance and inhibitory control differentiate from early childhood to adolescence. Developmental Neuropsychology. 2010; 35(6):679–697. [PubMed: 21038160]

Smith-Donald R, Raver CC, Hayes T, Richardson B. Preliminary construct and concurrent validity of the preschool self-regulation assessment (psra) for field-based research. Early Childhood Research Quarterly. 2007; 22(2):173–187.

Stuss, D.; Knight, R., editors. Principles of frontal lobe function. New York: Oxford; 2002.

Thorell LB, Wahlstedt C. Executive functioning deficits in relation to symptoms of adhd and/or odd in preschool children. Infant and Child Development. 2006; 15(5):503–518.

Toga AW, Thompson PM, Sowell ER. Mapping brain maturation. Trends in Neurosciences. 2006; 29(3):148–159. [PubMed: 16472876]

Welsh JA, Nix RL, Blair C, Bierman KL, Nelson KE. The development of cognitive skills and gains in academic school readiness for children from low-income families. Journal of Educational Psychology. 2010; 102(1):43–53. [PubMed: 20411025]

Wiebe SA, Espy KA, Charak D. Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. Developmental Psychology. 2008; 44(2):575–587. [PubMed: 18331145]

Willoughby M, Kupersmidt J, Voegler-Lee M, Bryant D. Contributions of hot and cool self-regulation to preschool disruptive behavior and academic achievement. Developmental Neuropsychology. 2011; 36(2):162–180. [PubMed: 21347919]

Willoughby MT, Blair CB. Test-retest reliability of a new executive function battery for sse in early childhood. Child Neuropsychology. 201110.1080/09297049.2011.554390

Willoughby MT, Kupersmidt JB, Voegler-Lee ME. Is preshcool executive function causally related to academic achievement? Child Neuropsychology. (in press).

Willoughby MT, Wirth RJ, Blair CB, Greenberg M. Investigators, F. L. P. The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. Psychological Assessment. 2010; 22(2):306–317. [PubMed: 20528058]

Woodcock, RW.; McGrew, KS.; Mather, N. Examiner’s manual. Woodcock-johnson iii tests of achievement. Itasca: Riverside Publishing; 2001.

Yung Y. Finite mixtures in confirmatory factor analysis models. Psychometrika. 1997; 62(3):297–330.

Zelazo, PD.; Müller, U. Executive function in typical and atypical development. In: Goswami, U., editor. Hanbook of childhood cognitive development. Blackwell; Oxford: 2002. p. 445-469.

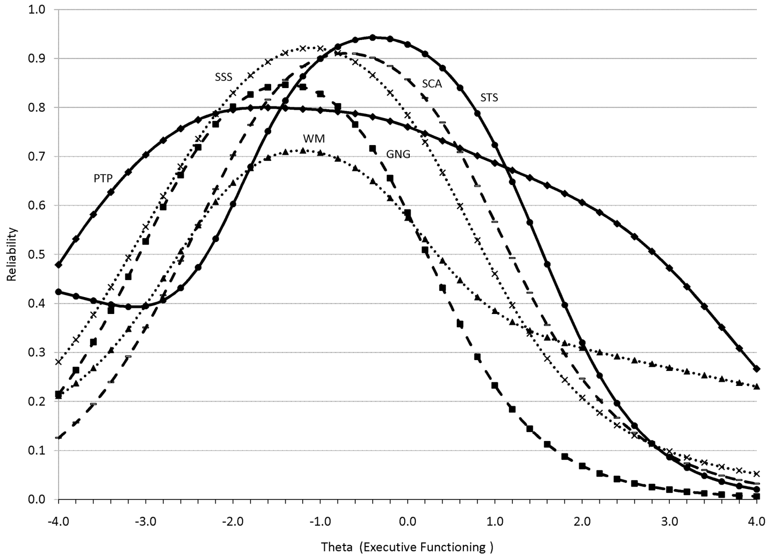

Figure 1.

Overlay of the Spatial Conflict Arrows (SCA), Silly Sound Stroop (SSS), Animal Go No-Go (GNG), Something is the Same (STS), Working Memory Span (WM), and Pick the Picture (PTP) reliability curves.

Table 1

Sample Description at Age 5 Year Visit

		EF Task Summary			
Descriptor		**Total (N=1091)**	**Complete (N=1036)**	**No-Opp (N=46)**	**Unable (N=9)**
		%	%	%	%
State	PA	40	42	9	33
Race	PC (AA)	42	41	43	44
	TC (AA)	43	43	48	44
Gender	PC (Female)	97	97	100	100
	TC (Female)	50	50	48	44
Income	Poverty at recruit	78	78	76	89
PC Education	4+ year degree	17	17	17	22
PC Marital status	Married	59	58	63	56
		M (*SD*)	*M*	*M*	*M*
Age	PC (Years)	31.6 (7.2)	31.6	31.9	33.8
	TC (Months)	60.6 (3.1)	60.4	65.0	60.1
PC	Education (years)	13.1 (2.0)	13.1	13.2	13.2
Household	Income/Needs	1.9 (1.6)	1.9	1.9	1.9

Note: PA=Pennsylvania; AA=African American; PC=primary caregiver; TC=target child; Completed = Children who completed one or more EF tasks; No-Opp = Children who were not given an opportunity to complete EF tasks; Unable = Children who were unable or unwilling to complete one or more EF tasks.

Table 2

Descriptive Statistics for and Bivariate Correlations between Executive Function Tasks and Academic Achievement Tests

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	***M***	***SD***	***Skew***	***Kurtosis***
1. SCA	--	.15	.14	.11	.06	.16	.22	.22	.24	.14	.19	0.0	0.9	−0.4	−0.7
2. SSG	.15	--	.28	.14	.18	.31	.23	.22	.22	.14	.25	−0.0	0.8	−0.8	−0.0
3. GNG	.12	.26	--	.15	.17	.29	.28	.23	.24	.20	.23	−0.0	0.8	−1.1	0.7
4. STS	.11	.15	.17	--	.14	.24	.32	.38	.37	.22	.26	0.0	0.9	−0.2	−0.9
5. WMS	.06	.18	.17	.12	--	.26	.20	.18	.20	.06	.16	−0.0	0.8	−0.6	−0.2
6. PTP	.15	.29	.29	.25	.26	--	.37	.35	.35	.24	.33	0.0	0.9	−0.7	1.0
7. ECLS-K Math	.18	.20	.26	.35	.20	.38	--	.68	.77	.63	.53	−1.7	0.7	0.2	−0.6
8. WJ - AP	.19	.22	.23	.41	.17	.37	.69	--	.73	.56	.56	100.4	12.5	−0.2	1.0
9. WJ - QC	.20	.21	.22	.39	.19	.37	.78	.73	--	.68	.55	91.9	13.8	0.5	−0.2
10. WJ - LW	.11	.13	.19	.22	.06	.25	.64	.58	.69	--	.45	98.4	13.3	−0.1	0.9
11. TOPEL - PA	.15	.23	.25	.29	.14	.34	.53	.54	.54	.45	--	92.9	14.4	−0.1	−0.2

Note: N = 1058; values above and below the diagonal are unweighted and weighted, respectively; SCA- Spatial Conflict Arrows; SSG -Silly Sound Stroop; GNG - Animal Go No-go; STS - Something's the Same; WMS - Working Memory Span; PTP - Pick the Picture; ECLS-K - Early childhood longitudinal study – kindergarten class; WJ - Woodcock Johnson; AP - Applied Problems; QC - Quantitative Concepts; LW - Letter-Word; PA - Phonological Awareness