Published in final edited form as:

Psychol Assess. 2012 June ; 24(2): 418–431. doi:10.1037/a0025779.

Executive Function in Early Childhood: Longitudinal Measurement Invariance and Developmental Change

Michael T. Willoughby¹, R.J. Wirth², Clancy B. Blair³, and Family Life Project Investigators

¹FPG Child Development Institute, University of North Carolina at Chapel Hill

²Vector Psychometric Group

³Department of Applied Psychology, New York University

Abstract

This study tested the longitudinal measurement invariance and developmental changes of a newly developed battery of executive function (EF) tasks for use in early childhood. The battery was administered in the Family Life Project—a prospective longitudinal study (N = 1292) of families who were over-sampled from low income and African American families at the birth of a new child—at age 3, 4, and 5 year assessments. All six individual EF tasks exhibited strong measurement invariance over time. The EF battery, which was derived from the six individual tasks, exhibited partial strong invariance over time. Second-order latent growth curve models revealed individual differences in the levels but not rates of change in latent EF ability. The functional form of change was nonlinear; 60% of the total change in EF ability that was observed between the 3-5 year assessments occurred between the year 3 and 4 assessments. Results are discussed with respect to the importance of establishing scalable measures of EF ability prior to investigating experiences that predict and/or are predicted by changes in EF during early childhood.

Keywords

Executive Function; Early Childhood; Measurement Invariance; Latent Growth Curve Models

Executive functions (EF) refer to cognitive abilities 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 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 *mental flexibility*, defined as the ability to shift attentional or cognitive set among distinct but related dimensions or aspects of a given task (Davidson *et al.*, 2006; Diamond *et al.*, 2002; Garon *et al.*, 2008; Zelazo & Müller, 2002).

Over the past two decades, a number of investigators have demonstrated the feasibility of measuring EF in young children (Diamond & Taylor, 1996; Espy, 1997; Zelazo & Reznick,

1991) and that EF is meaningfully related to a number of aspects of child development including social-emotional competence (Carlson, Mandell & Williams, 2004; Hughes & Ensor, 2004) and early academic ability (Blair & Razza, 2007; Bull & Scerif, 2001; Espy et al., 2004). Although a number of studies have examined task conditions under which children will or will not exhibit executive cognitive abilities (Diamond, Carlson, & Beck, 2005; Zelazo, Muller, Frye, & Marcovitch, 2003), measures of EF abilities in young children have for the most part have been appropriate for use only at single time points, demonstrating sufficient variability in performance at relatively narrow age ranges. To address this aspect of EF measurement and to facilitate the investigation of the development of EF across the early childhood period, we developed a battery of EF tasks for use in large-scale longitudinal studies. Specifically, we set out to develop a task battery that was highly portable, that presented a variety of tasks in a uniform format, that was easily administered by lay staff, that elicited individual differences in ability level, and that resulted in scores that were scalable across the preschool period (i.e., age 3-5 years), thereby permitting the analysis of longitudinal (within-person) change.

Thus far, we have considered the psychometric properties of the new battery in a cross-sectional convenience sample involving 3-5 years-olds, as well as at individual time points, age 3, 4, and 5-years, in the large, representative sample of children analyzed longitudinally here (Willoughby & Blair, 2011; Willoughby *et al.*, 2011a; Willoughby *et al.*, 2011b; Willoughby *et al.*, 2010). Three primary conclusions have resulted from this previous work. First, children's performance on our task battery was best represented by a single latent factor. Although we selected EF tasks that represented the tripartite organization of EF, empirically, we found no evidence for this differentiation. Our results converge with other recent studies, which also demonstrated that children's performance on tasks that purportedly assess inhibitory control, working memory, and attention shifting processes, is best conceptualized as unidimensional (Espy *et al.*, 2010; Hughes *et al.*, 2010; Shing *et al.*, 2010; Wiebe *et al.*, 2008). Second, whereas the short-term (approximately 2 to 3 weeks) test-retest reliability of individual tasks in our battery was poor to modest, $r_s \approx .60$, the test-retest reliability of the battery was excellent when examined using a latent variable model, $\phi = .95$. This result underscores an under-appreciated finding from studies that have utilized confirmatory factor analyses (CFA) of EF tasks, both with adult and child samples; namely, only one-fourth to one-half of the observed variation in any given EF task represents true ability (i.e., the R^2 estimates of EF tasks that serve as indicators of an EF latent variable range from .25 - .50). Aggregating performance across tasks improves the precision of measurement of EF ability and is more preferable than the analysis of individual tasks in isolation. Third, our task battery does a relatively better job of measuring EF ability in the low to average proficiency range than it does in measuring ability in the extremely high or low proficiency range. This stems from our initial task development efforts, which prioritized the construction of tasks that were amenable for use with young children (especially three-year-olds) who resided in low-income homes. The primary objectives of this current study were (1) to extend our previous psychometric work by testing for longitudinal measurement invariance of our individual tasks, as well as the task battery scores and (2) to describe the developmental trajectory of EF abilities as measured by our task battery from age 3 through 5 years in a large, representative sample.

Executive functions are subserved by multiple neural circuits involving interconnections of prefrontal cortex (Miller & Cohen, 2001; Stuss & Knight, 2002) with striatal (Durstun & Casey, 2006; Durstun *et al.*, 2002; Liston *et al.*, 2006) and parietal regions (Edin et al., 2009; Klingberg, 2006). Given the protracted developmental course of prefrontal cortex (Toga *et al.*, 2006), precise longitudinal measurement of executive function abilities in children has been a priority. Numerous cross-sectional studies, which compared groups of selectively sampled individuals of different ages, have demonstrated that EF abilities develop

throughout childhood and well into adolescence and early adulthood (e.g., Chelune & Baer, 1986; Davidson et al., 2006; Huizinga *et al.*, 2006; Levin *et al.*, 1991; Luna *et al.*, 2004; e.g., Rebok *et al.*, 1996; Welsh *et al.*, 1991). Romine and Reynolds (2005) provided a quantitative summary of eight such studies, which demonstrated relatively large changes in EF during middle childhood that appeared to begin to slow during late adolescence. However, as they noted, the specific functional form, including the age at which an asymptote is reached, is difficult to discern due to uncertainties in the psychometric properties of commonly used tasks. Consistent with this observation, Davidson and colleagues (2006) demonstrated how an apparent asymptote in EF performance that is observed in adolescence may be an artifact of task scoring (i.e., although accuracy scores appear to asymptote, reaction time indices of ability do not). The important point is EF tasks (and batteries) that exhibit strong psychometric properties will provide the best opportunity to delineate the functional form of EF change across distinct periods of development.

The current study focused on developmental changes in EF abilities in children age 3-5 years. Early childhood is understood to be a period of pronounced developmental improvements in EF abilities (Anderson *et al.*, 2008; Carlson, 2005). Empirical support for this assumption comes principally from studies that used between subjects, cross-sectional designs to demonstrate that older preschool-aged children outperform younger preschool-aged children on a variety of EF tasks (e.g., Diamond et al., 2002; Dowsett & Livesey, 2000; Gerstadt *et al.*, 1994; Jones *et al.*, 2003; Luciana & Nelson, 1998; Smidts *et al.*, 2004). Secondary evidence comes from studies that have employed electroencephalograph (EEG) methods. These studies have documented changes in prefrontal-based circuitry during the early childhood period (Marshall *et al.*, 2002; Thatcher *et al.*, 2008), as well as evidence that individual differences in EEG power and/or coherence during the completion of EF tasks are associated with enhanced EF performance (Bell & Wolfe, 2007; Swingler *et al.*, in press).

To the best of our knowledge, Hughes, Ensor and colleagues (2011, 2010) were the first (and only) group to investigate developmental changes in EF abilities among typically developing preschool-aged children using a within subjects, prospective longitudinal design (note that Diamond and colleagues (1997) used a prospective longitudinal design when investigating developmental changes in EF abilities among children treated for PKU). The advantages of using prospective longitudinal designs to inform developmental changes in EF ability are numerous. Foremost among them is the ability to partition between and within sources of variance in EF scores (i.e., consideration of inter-individual differences in intra-individual change), the ability to more confidently delineate the functional form of change (particularly if the form is nonlinear), and the ability to use growth parameters (e.g., intercepts, slopes) as both independent and dependent variables. One of the strengths of the Hughes and Ensor's (2010) work was the establishment of the longitudinal measurement invariance of their three-task battery across time (i.e., at ages 4 and 6 years). They demonstrated that EF tasks worked, in a psychometric sense, equally well at age 4 and 6 year assessments. Unfortunately, their results did not inform questions about the magnitude of changes in EF abilities that occurred between ages 4 and 6 years. To be clear, it is our supposition that Hughes and Ensor's latent growth curve models were incorrectly parameterized, which limited the ability to draw any meaningful inferences about the magnitude of change in EF abilities between 4 and 6 years of age.

The current study represents the second study to use a prospective longitudinal design to inform developmental changes in EF ability during early childhood. In addition, the current study sought to build on the specific strengths of two previous studies that investigated age-based changes in EF ability involving older children. Luna and colleagues (2004) were unique in that they paid special attention to understanding the functional form of change in EF across age. In contrast to the widely used polynomial parameterizations of change, they

reported that inverse regression models provided the best fit to their data. The consideration of specific nonlinear functions of change in EF may facilitate more precise thinking about the developmental processes that account for this change (Ram & Grimm, 2007). Rather than characterizing developmental changes on *specific* EF tasks, Huizinga and colleagues (2006) were unique in their focus on latent EF ability (estimates of “true” ability). Latent variable methods provide a principled method for understanding changes in EF ability that are free of measurement error and/or systematic variation that is task specific. Building on the strengths of the studies by Luna and Huizinga, the current study utilized second-order latent growth curve (LGC) models to evaluate inter-individual differences in intra-individual change of latent EF ability between 3-5 years of age (Bollen & Curran, 2006; Ferrer *et al.*, 2008). Unlike the more commonly used first-order LGCs, which characterize changes in a *measured outcome* variable across time, second-order LGCs simultaneously estimate latent variables, which represent *latent variable score* variation that is shared across multiple measured outcome variables at each assessment occasion, and characterize changes in this latent variable score variation across time. Moreover, we parameterized second-order LGCs in a way that facilitate consideration of whether developmental changes in EF ability were linear versus nonlinear across time.

In sum, the current study tested the longitudinal measurement invariance of individual EF tasks, as well as of a battery-derived score. Tests of measurement invariance placed task and battery derived scores on a common developmental metric. Second-order LGC models were used to test the functional form of change in latent variable scores of EF ability measured prospectively from ages 3-5 years.

Methods

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, 1999). 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). Full details of the sampling procedure appear elsewhere (Vernon-Feagans, Cox, and the Family Life Key Investigators, 2011).

Of those families interested and eligible and selected to participate in the study, 1292 families completed a home visit at 2 months of child age, at which point they were formally enrolled in the study. A total of N=1123 (87%), N=1066 (83%), and N=1099 (85%) families participated in the age 3-, 4-, and 5-year assessments, respectively. While most families (N=1017; 79%) participated in all three assessments, some participated in two (N=86; 7%) or one (N=65; 5%) of the three possible assessments. Approximately 10% (N=124) families who were enrolled in the study at the 2-month visit did not participate in any of the 3-, 4-, or 5-year assessments. Families and children who did not participate in any of the 3-, 4-, or 5-year assessments (N = 124), and hence who were not included in this study, did not differ from families who did participate (N=1168) with respect to child race (38% vs. 43% African American, $p = .28$), child gender (56% vs. 50% male, $p = .19$), state of residence (42% vs. 40% residing in PA, respectively, $p = .67$), or being recruited in the low income stratum (76% vs. 78% poor, $p = .62$).

Procedures

Families participated in home visits when study children were 3 (2 visits), 4 (1 visit) and 5 (1 visit) years old. Home visits consisted of a variety of parent (cognitive testing, interviews, questionnaires) and child (including EF) tasks; the 3 and 5 year assessments also included dyadic (parent child interactions and book-reading) tasks that were videotaped for later coding. On average, each home visit took approximately two hours to complete, of which children spent approximately 30-45 minutes (including practice, administration, and breaks between tasks) completing EF tasks. One research assistant (RA) was responsible for administering EF tasks (in a fixed order) to children, including keeping children 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. Neither RA was responsible for evaluating the accuracy of child responses. Computerized scoring, which took place when data for the entire visit were processed, was used to evaluate the accuracy of child responses to each task. This scored, item-level data formed the basis of psychometric analyses included herein.

Measures

Executive function—The set of EF tasks shared a number of features. Each task was presented in an open spiral bound flipbook format (pages measured 8" × 14") which allowed the examiner to easily turn pages that present stimuli on one page and scripted instructions for administration on the other. For each task, examiners 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 testing on that task. At the conclusion of each task, the RA who recorded child responses made a 3-point rating regarding the quality of the data for that task (i.e., 1 = low quality, indicated by frequent distractions or concerns about the child's ability to understand what is being asked of him/her; 2 = average quality, indicated by few distractions and confidence that the child understands what is being asked of him/her; 3 = high quality, indicated by no distractions and exceptional effort or engagement of the part of the child). Ratings were intended to convey how amenable the household was for purposes of testing, as well as child apparent comprehension and engagement in the task—not his/her performance. Brief descriptions of each task are provided below. Elaborated task descriptions, which include sample pictures of test stimuli, are available by request from the first author.

Working Memory Span (WMS; Working Memory)—This span-like task is based upon principles described by Engle, Kane and collaborators (e.g., Kane & Engle, 2003). In this task, children are presented with a line drawing of an animal figure above which is a 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 trials, two 3-house trials, and two 4-house trials. Responses were summarized as the number of items answered correctly within each item set. Although the WMS was administered at all three assessments, only the first 11 items were administered at the 3-year assessment (i.e., based on the results of pilot testing and test burden, the 4-house trials were omitted), whereas all 19 items were administered at the 4 and 5-year assessments.

Pick the Picture Game (PTP; Working Memory)—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 of an apple and dog. For the first page, they pick (touch) either of the two pictures. For the second page they are requested to pick a different picture. The arrangement of pictures within each set is randomly changed across trials (including some trials not changing) 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. Children received two each of 2-picture, 3-picture, 4-picture, and 6-picture sets (the first item in each set was not scoreable and only used to define the accuracy of the remaining responses in each picture set). Responses were summarized as the number of items consecutively answered correctly in each picture set beginning with the first item in the set. Through pilot testing, it was determined that the PTP was too difficult for many 3 year olds; hence, it was only administered at the 48- and 58-month assessments.

Silly Sounds Stroop (SSS; Inhibitory Control)—This task was derived from the Day-Night task developed by Gerstadt, Hong, and Diamond (1994). Children are presented with pictures of a cat and dog. The experimenter asks the child to make the sounds of a dog and then a cat. The experimenter then introduces the idea that, in the Silly Sounds game, dogs make the sounds of cats and vice versa. Scripted coaching and elaboration is provided. Then pages of a flip book are presented that contain side-by-side pictures of cats and dogs (in random order). The experimenter points to the first picture and asks what sound this animal makes in the Silly Sounds game and then points to the adjacent picture and asks the same question. In terms of administration, verbal prompts (i.e., “what sound does this animal make in the Silly Sounds game?”) are discontinued after the first 8 items (the experimenter just flips a page and points to pictures). A total of 36 items are presented (18 flip book pages). Responses (correct, incorrect) to the first item on each page were used for purposes of scoring. The SSS was administered at all three assessments.

Spatial Conflict (SC; Inhibitory Control)—This is a Simon task similar to that used by Gerardi-Caulton (2000). A response card, which has a picture of a car on the left side and picture of a boat on the right side, is placed in front of the child. The RA turned flip book pages that depict either a car or boat. The child was instructed to touch the car on his/her response card when the flip book page showed a car and to touch the boat on the response card when the page showed a boat. Across the first 8 trials, cars and boat were depicted centrally. These items provide an opportunity to teach the child the task (“touch your car when you see a car, touch your boat when you see a boat”). For items 9-22, cars and boat are depicted laterally, with cars always appearing on the left side of the flip book page (“above” the car on the response card) and boats always appearing on the right side of the flip book page (“above” the boat on the response card). Children are instructed to touch the “car” with their left hand (left hand touches left presented stimuli) and to touch the “boat” with their right hand (right hand touches right presented stimuli). For items 23-35, cars and boat begin to be depicted contra-laterally, with cars usually (though not exclusively) appearing on the right side of the flip book page (“above” the boat on the response card) and boats appearing on the left side of the flip book page (“above” the car on 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). Responses (correct, incorrect) to contra-laterally presented items were used for purposes of scoring. The SC was administered at the 3-year assessment.

Spatial Conflict Arrows (SCA; Inhibitory Control)—This task was identical in format to the SC task described above with the exception that the response card consisted of two buttons (black circles) and the stimuli were arrows that pointed to either the left or right. The child was instructed to touch the left button when arrows pointed to the left and to touch the right button when arrows pointed to the right. As with the SC task, across the first 8 trials, arrows were depicted centrally. These items provided an opportunity to teach the child the task (“touch this [left] button when arrows points this [left] way and touch this [right] button when arrows point this [right] way”). For items 9-22, arrows were depicted laterally, with left pointing arrows always appearing on the left side of the flip book page (“above” the left button on the response card) and right pointing arrows always appearing on the right side of the flip book page (“above” the right button on the response card). As before, children were encouraged and prompted (but not required) to touch the left button with their left hand and right button with their right hand. For items 23-35, arrows began to be depicted contra-laterally, with left pointing arrows appearing on the right side of the flip book page and right pointing arrows appearing on the left side of the flip book page. 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). Responses (correct, incorrect) to contra-laterally presented items were used for purposes of scoring. The SCA was administered at the 4 and 5-year assessments.

Animal Go No-Go (GNG; Inhibitory Control)—This is a standard go no-go task (e.g., Durston *et al.*, 2002) presented in a flipbook format. Children are presented with a large button that clicks when pressed. They 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. Responses (correct, incorrect) to no-go trials were used for purposes of scoring. The GNG was administered at all three assessments.

Something’s the Same Game (STS; Attention Shifting)—This task was derived from Jacques and Zelazo’s (2001) flexible item selection task. In this task, children are shown a page containing two pictures that are similar along one dimension (content, color, or size). The experimenter then explicitly states the dimension of similarity. The next page presents the same two pictures, plus a new third picture. The third picture is similar to one of the first two pictures along a dimension that is different from that of the similarity of the first two pictures (e.g., if the first two pictures were similar along the dimension of shape, the third card would be similar to one of the first two along the dimension of color or size.) Children are asked to choose which of the two original pictures is the same as the new picture. This requires the child to shift his/her attention from the initial dimension of similarity to a new dimension of similarity. The person scoring the task only records which picture the child touched on each trial. Twenty trials are presented (only the first 15 items were presented at the 3-year assessment. Responses (correct, incorrect) to all but the first item are used for scoring (the first item is excluded from scoring because incorrect answers are corrected in order to teach the task). The STS was administered at all three assessments.

Analytic Strategy

Analyses, which proceeded in four stages, tested the longitudinal measurement invariance of individual EF tasks, the longitudinal measurement invariance of the overall EF battery, and the functional form of changes across assessments (age 3-5 years). In the first stage, confirmatory factor analyses (CFA) were used to evaluate the longitudinal measurement invariance of all six individual EF tasks. This resulted in a single longitudinally scalable

score for each child on each EF task that was completed. In the second stage, item response theory (IRT) models were used to score each task, as well as to evaluate reliability curves for each task. In the third stage, CFAs were used to evaluate the longitudinal measurement invariance of the task battery (where children's performance on each task, scored in the previous step, served as indicators of underlying EF ability). This resulted in a single longitudinally scalable score for each child who completed at least one task at any given assessment occasion. In the fourth stage, a series of second-order latent growth curve (LGC) models were estimated in order to test the functional form and degree of change in EF ability from age 3- to 5-years. Given only three measurement occasions, we only considered linear and free-loading models (Bollen & Curran, 2006; Meredith & Tisak, 1990). These models differed in the factor loading structure that was imposed on the data. Whereas the linear model assumed a constant rate of change between adjacent assessments (i.e., factor loadings were fixed to values of -11, 0, 12, representing the mean number of months, based on average child age, that transpired between assessments; age was centered around the age 4-year assessment), the freed loading model estimated the first (or equivalently last) factor loading allowing the model to "stretch" time to better fit the observed data (the other two factor loadings took on values identical to those used in the linear model parameterization). To the extent that estimated factor loading differed from those imposed in the linear model, this would indicate nonlinear change is evident. Because the freed loading model is nested in the linear model, a chi-square difference test was used to determine which functional form best represented change in latent variable scores of EF ability across time.

IRT models used Bock-Aitkin EM estimation as implemented in IRTPRO (Cai, du Toit, & Thissen, forthcoming). IRT models included either 2-parameter logistic models (2PLMs) for those tasks with dichotomous items response formats (i.e., SC, SCA, SSS, GnG, STS) or graded responses models (GRMs) for those tasks with polytomous item response formats (i.e., PTP, WMS). Consistent with earlier reports, two of the tasks involved bi-factor IRT models (i.e., the 2PLM for SSS contained separate method factors for 'cat' and 'dog' items; the STS scale contained method factors for size, color, and object items). CFA and LGC models used robust maximum likelihood (MLR) for models involving continuous indicators and weighted least squares with mean and variance adjustment (WLSMV) for models involving dichotomous and ordinal indicators—all as implemented in version 5.2 of *Mplus* software (Muthén & Muthén, 1998-2007). Following conventions proposed by Hu and Bentler (1999), CFA and LGC models that had comparative fit indices (CFIs) $\geq .95$ and root mean squared errors of approximation (RMSEA) $\leq .05$ were considered to exhibit good fit, while those models with CFI $\geq .95$ or RMSEA $\leq .05$ were considered to exhibit adequate fit. Item-level analyses (IRT and CFA models of individual tasks) did not take into account the complex sampling design (stratification and over-sampling) but battery-level analyses (CFA and LGC models) did.

Results

Participation in EF Assessments

As we have described in companion manuscripts, a small number of children whose families participated at each assessment did not complete EF assessments. In some cases, families had moved out of the study area (defined as a 200 mile radius), in which case their visit was conducted by telephone, which prohibited child EF assessments. Among the majority of families who completed home visits, up to 9% of children at any given assessment were unable or unwilling to complete EF tasks. This was more common at earlier than later visits (e.g., 91% vs. 99% completion rates at the 3 and 5-year assessments, respectively). As detailed elsewhere, children who were unable to complete tasks typically had behavioral, cognitive and/or physical impairments that prohibited their participation (Willoughby et al., 2010).

In total, $N=973$ (75% of the total sample), $N=1009$ (78%), and $N=1038$ (80%) children participated in EF testing at 3, 4, and 5-year home visits, respectively. A total of $N=865$ (67% of the total sample) participated in EF testing at all three assessments, $N = 169$ (13%) at 2/3 assessments, $N = 87$ (7%) and 1/3 assessments, and $N = 171$ (13%) at none of the assessments. Children who were enrolled in the study but who did not complete any EF assessments (typically but not exclusively due to family non-participation in these assessments; $N = 171$) did not differ from children who did participate in at least one EF assessment ($N=1121$) with respect to child race (37% vs. 43% African American, $p = .15$), child gender (56% vs. 50% male, $p = .19$), state of residence (36% vs. 41% residing in PA, respectively, $p = .26$), or being recruited in the low income stratum (77% vs. 78% poor, $p = .75$).

The $N = 973$ children who participated in EF testing at the 3-year assessment were on average 38 ($SD = 2.0$; Median 37.4) months old and completed an average of 3.5 ($SD = 1.3$, Median = 4.0) of the 5 EF tasks. The $N = 1009$ children who participated in EF testing at the 4-year assessment were on average 49 ($SD = 1.6$; Median 48.7) months old and completed an average of 5.5 ($SD = 1.0$, Median = 6.0) of the 6 EF tasks. The $N = 1038$ children who participated in EF testing at the 5-year assessment were on average 61.3 ($SD = 3.1$; Median 62.4) months old and completed an average of 5.8 ($SD = 0.7$, Median = 6.0) of the possible 6 tasks.

Individual Tasks: Longitudinal Measurement Invariance, Scoring, & Reliability Curves

We have previously established that a single factor model adequately explained the covariation of items on each task, though some tasks also require method factors, which account for residual covariances between subsets of items. Here, we tested whether each of the six EF tasks exhibited longitudinal invariance across all three assessments. A separate longitudinal CFA model was fit to each task, constraining each item's factor loading and threshold parameters to be equal across assessments. For purposes of identification, the mean and variance for the general factor at the 4-year assessment were constrained to zero and one, respectively. All item-level unique variances, the latent mean and variances for the 3 and 5-year general factors, and all of the correlations between latent EF factors were freely estimated. For bi-factor models (i.e., those models involving method factors), correlations between latent EF factors and method factors as well as correlations between method factors were constrained to zero. As summarized in Table 1, results indicated that each task met the conditions of strong measurement invariance. Longitudinal CFAs demonstrated acceptable to good levels of fit, indicating that item parameters (thresholds, factor loadings) could take on identical values for each tasks across all three assessments without resulting in poor model fit. Practically, this indicated that individual EF tasks "worked", in a psychometric sense, equivalently at the age 3, 4, and 5-year assessments.

The next step involved creating IRT-based (i.e., expected a-posteriori [EAP]) scores for each task that was completed by each child at each assessment. We have previously elaborated both the process and merits of using EAPs to score each EF task, using select data from the 48-month assessment as an exemplar (Willoughby et al., 2011a). Here, we extend that scoring approach to the longitudinal setting (i.e., we provide EAPs for each task that are on a common developmental scale spanning 3, 4, and 5-year assessments). This was accomplished through the use of a calibration sample. Specifically, for each task, a random sample of children was drawn across the 3, 4, and 5-year assessments, resulting in a sample of children who completed the task at 3, 4, or 5-year assessments (no child contributed data from more than one assessment). A calibrated sample was established by randomly selecting one time of assessment from children who had completed four of the five tasks at the 3-year assessment or five of the six tasks at 4 or 5-year assessments *and* who were rated (by RAs) as average or above average quality (task specific $Ns = 929 - 1045$). Selecting children who

performed well on the tasks (i.e., as evidenced by both the total number of tasks completed at a given assessment occasion and RA ratings regarding testing conditions and child effort—but not accuracy) was intended to improve the quality of data that was used to estimate parameters that informed task scoring. Calibration samples represent a commonly used strategy in situations where IRT models are used to inform longitudinal scoring (e.g., Hussong, et al, 2007, Curran, et al, 2008). In our case, the use of a calibration sample substantially reduced the complexity of the models to be estimated. Rather than attempting to simultaneously estimate a common set of IRT parameters for each item of each task simultaneously for 3, 4, and 5-year assessments, only a single set of item parameters was estimated for the calibrated subsample of children who were drawn from 3, 4, and 5-year assessments.

Descriptive statistics for EAP scores for each task at each assessment are summarized in Table 2 (scores are interpreted as having a z score metric, where average performance is defined at the sample mean age across assessments—approximately 50 months). As expected, children's performance on all tasks increased from 3-5 year assessments. Whereas some tasks appeared to exhibit relative constant (linear) change over time (e.g., SSS), others appeared better characterized by nonlinear change (e.g., STS).

One of the benefits of adopting an IRT-based approach for task evaluation and scoring is the ability to compute reliability curves. Reliability curves characterize the precision of measurement (i.e., changes in the standard error of measurement) of each task as a function of child ability level measured free of error. Figure 1 depicts reliability curves for all seven EF tasks (i.e., reliability is plotted as function of latent EF ability level, which is referred to as “theta” in IRT parlance). In general, tasks did a relatively better job of measuring EF for children whose true ability was in the low to average (versus average to high) range. One strategy for improving the measurement of EF ability, particularly for children with very low or high ability levels, is to aggregate information about children's performance across multiple tasks.

Task Battery: Longitudinal Measurement Invariance

We have previously established that children's performance on the EF task battery was best represented by a single latent factor at separate assessments. Here, we tested whether individual EF tasks conformed to a single factor in an equivalent way across time (i.e., whether the contribution of individual tasks to the underlying latent factor representing EF ability changed across assessments). Children's performance (i.e., EAPs scores) on five tasks (i.e., SC, SSS, GNG, STS, WMS) served as indicators of underlying EF ability at the 3-year assessment, while performance on six tasks (i.e., SCA, SSS, GNG, STS, WMS, PTP) served as indicators of EF ability at the 4 and 5-year assessments. Four tasks that were administered to children at all three assessments were used to define a common scale for latent EF ability across assessments (i.e., SSS, GNG, STS, WMS). The SC(A) and PTP tasks served as supplemental indicators of latent EF ability across 3, 4, and 5-year assessments.

A longitudinal CFA model was estimated which imposed strong longitudinal invariance. This model equated factor loadings and intercepts for any task that was repeated across assessment. This model only fit the data moderately well, $\chi^2(116) = 252.5, p < 0.001$, CFI = 0.89, RMSEA = 0.03. In order to determine whether some tasks were better indicators of EF ability at different assessment periods, we examined residual matrices, as well as results from CFA models that considered constraints for each task individually. This resulted in the determination of a sequence of CFA models, which represented differing degrees of partial longitudinal measurement invariance, to be tested using nested chi-square tests. As summarized in Table 3, a longitudinal CFA model that constrained the factor loadings and item intercepts for SSS and PTP tasks, but allowed other tasks to have differential

contributions to the EF ability, provided an excellent fit to the data, $\chi^2(102) = 145.4$, $p = 0.003$, CFI = 0.97, RMSEA = 0.02.

Considering results from the final CFA model, all of the tasks had significant factor loadings at each assessment; however, standardized factor loadings ranged from .20 - .60 (task $R^2 = .04 - .36$), consistent with only modest inter-correlations between children's performance on individual tasks at each assessment. Children's performance on the battery was highly correlated across time (latent correlations of EF ability: ϕ_3 & 4 years = .88, ϕ_3 & 5 years = .86, ϕ_4 & 5 years = .91, $p < .001$, respectively). Collectively, these results suggest that whereas individual EF tasks are "noisy" indicators of true (latent) ability level, aggregating children's performance across tasks results in a highly stable index of true (latent) ability level between 3 and 5 years of age.

Task Battery: Second Order Latent Growth Curve Models

As a final step, we estimated a series of second-order LGC models which characterized changes in latent EF ability across 3, 4, and 5-year assessments. In order to place growth parameters on an interpretable metric, the mean and variance of the latent intercept were fixed to values of 0 and 1 and time was centered around the age 4-year assessment (i.e., the factor loading for the 4-year assessment was fixed to 0 in both linear and freed loading models). This resulted in growth parameters that were interpretable in the z-score metric used to define EAPS. The second-order *linear* LGC model fit the data well, $\chi^2(105) = 163.2$, $p = .0002$, CFI = 0.95, RMSEA = 0.02. The linear slope term was statistically significant ($\mu_\beta = 0.12$, $p < .001$), indicating model-implied improvements of 12 tenths of a standard deviation in children's latent variable score EF ability level with each passing month between ages 3 and 5 years. The latent variance for the slope parameter was not statistically significant ($p = .58$). That is, this model did not differentiate inter-individual differences in intra-individual change in true EF ability level. The inability to detect individual differences in rates of change in performance across time explained the non-significant correlation between growth parameters ($\phi = .49$, $p = .43$).

The second-order *freed loading* LGC model also fit the data well, $\chi^2(104) = 151.6$, $p = .002$, CFI = 0.96, RMSEA = 0.02. The estimated factor loading for the 3-year assessment was $\Lambda = -18.2$, $p < .001$ —compared to a fixed value of -12 in the linear LGC. Using an equivalent parameterization, the estimated factor loading for the 5-year assessment resulted in $\Lambda = 7.3$, $p < .001$ —compared to a fixed value of 11 in the linear LGC. A chi square difference test, which appropriately took into account the scaled nature of chi square test statistics resulting from the use of the MLR estimator, determined that the freed loading second-order LGC model fit the data better than the linear second-order LGC model, $\chi^2(1) = 9.1$, $p = .003$. The pattern of fixed (from the linear LGC) and estimated (from the freed loading LGC) factor loadings indicated that 60% of the total change in EF ability that was observed between the 3-5 year assessments occurred between the 3 and 4-year assessments (i.e., the freed loading models "stretched" the time scale to reflect greater rates of change between the first two versus the latter two assessments).

Figure 2 presents unstandardized coefficients for the second order LGC. Given our use of longitudinally-scaled EAP scores as indicators of latent variables and our parameterization of the model such that the intercept approximated the average age of the calibration sample, unstandardized estimates are more interpretable than are standardized estimates (i.e., growth parameters are interpretable on a z-score metric). For presentation purposes, residual (co)variances are not presented; however, R^2 values are. Whereas only between 4-38% of the observed variation in task (EAP) scores were explained by latent ability factors, 89-91% of the observed variation in the battery scores were explained by growth parameters. These results indicate that whereas individual EF tasks are noisy indicators of true ability,

individual differences in latent ability level follow orderly increases across time. Residual correlations for a given task across assessments (not presented) were of modest magnitude ($|r| = .00 - .20$) and often not statistically significant.

The freed loading model provided a statistically superior fit to the linear model. However, as is shown in Figure 3, both linear and freed loading (nonlinear) models provide similar representations of observed battery scores. Figure 3 also demonstrates marked differences in the variation in EF scores at each assessment. When scores are put onto a scalable metric across time, the *relative* variation in EF ability is smaller at the 3-year than at the 4 or 5-year assessments. While the variation at 3 year assessment appears “compressed” due to the longitudinal scale, there are still meaningful individual differences in mean level at all three assessments. Although we fixed the latent variance of the intercept parameter in the second order LGC models to 1 in order to facilitate interpretation, in models in which this parameter was freely estimated (not presented), it had a statistically significant variance. Hence, although the second-order models were unable to partition individual differences in rates of change in EF across time, they were able to partition individual differences in levels of EF ability at all three assessments.

Discussion

Given the relation of executive functions (EF) to a number of aspects of child development—including self-regulation, mental development, school readiness, and risk for psychopathology—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 factors that promote competence in children at risk for school failure and early developing psychopathology (Blair, Zelazo, & Greenberg, 2005). This study tested the longitudinal psychometric properties of a newly developed EF task battery, which represented an adaptation and extension of commonly used tasks in the literature. Results demonstrated that individual tasks exhibited strong measurement invariance across early childhood. The task battery exhibited partial strong measurement invariance across early childhood. Second-order LGCs demonstrated appreciable improvement in *latent* EF ability across early childhood, with indication of slightly faster improvement in latent ability level between 3–4 years relative to 4–5 years of age. Although there were stable inter-individual differences in EF ability at each assessment, there was no evidence for individual differences in the rate of change in EF across time.

All six individual EF tasks exhibited strong longitudinal measurement invariance. This ensures that any observed differences in mean ability on tasks are not due to differential measurement properties across time. Relatively few studies involving preschool-aged children have tested for measurement invariance of tasks across different age groups, with Wiebe and colleagues (2008) being an important exception, and even fewer have tested for *longitudinal* measurement invariance of the same children and tasks across time, with Hughes and Ensor (2010) being an important exception. Our results add to a growing body of work indicating that the measurement properties of individual tasks appear to be relatively robust between 3–5 years of age.

The reliability curves associated with each individual EF task underscore the point that even EF tasks that purportedly measure the same latent construct often do so with appreciably different precision of measurement, which varies as a function of children’s true ability level. All six tasks had reliability $\geq .75$ for children whose latent (true, error free) ability was in the low-average range (i.e., theta ranging from -1.2 to 0 , which is interpretable as child true-score ability on a z score metric). However, reliability curves were asymmetric;

all tasks did a *relatively* better job of measuring lower than higher levels of EF ability level. Reliability curves reveal that, *in their current form*, our individual tasks are not optimal for measuring EF ability in children with high ability levels (including precocious 4-5 year-olds and/or children older than 5 years). Reliability curves are directly related to item difficulty and discrimination parameters. The reliability curves in Figure 1 reveal that our tasks consist primarily of relatively easy items that have reasonably strong discrimination. As described below, we are currently modifying tasks to better measure a broader range of ability level (with a special emphasis on adding more difficult items). Ideally, this effort will result in the improved precision of measurement of EF latent (true) ability across a wider range of functioning. In general, more widespread presentation of reliability curves for EF tasks would facilitate researchers (and clinicians) ability to select tasks that optimally measure ability levels of interest. In the absence of individual EF tasks that have high levels of reliability across the full range of child EF ability, it is advantageous to aggregate children's performance across tasks using latent variable methods.

Longitudinal confirmatory factor analyses, in which children's performance on individual tasks at each assessment served as indicators of latent EF ability at each assessment, demonstrated that the task battery exhibited partial measurement invariance. Whereas the SSS and PTP tasks indexed EF ability in a comparable way across assessments, other tasks had factor loadings that varied in the strength of their association with EF ability across assessments. The ability to constrain the measurement properties of two tasks (SSS—across 3, 4, and 5-year assessments; PTP—across 4 and 5-year assessments) to take on equal values across time facilitated the creation of scalable battery scores, thereby permitting longitudinal analyses. The results of longitudinal CFAs also highlighted that although children's performance across tasks at any given assessment period was modest, their performance on the battery was highly stable over time.

An open question is how few of the individual EF tasks are necessary to be administered in order to obtain a reliable estimate of latent EF ability, as well as whether the selection of specific tasks should depend on child age or expected ability level. Although these questions were beyond the scope of this study, the reliability curves and the results from the longitudinal confirmatory factor analyses are informative. To the extent that a proposed study intended to use a prospective longitudinal design in order to focus on within person change, administration of the SSS and/or PTP tasks would appear prudent, given evidence that these are equally good indicators of EF ability from 3-5 and 4-5 years, respectively. The reliability curves assist in informing the choice of additional specific tasks. For example, a study that would primarily involve young children may benefit from the use of SC or SCA tasks, which have precise measurement of ability over a relatively narrow range of low to average ability. In contrast, a study that included a broader range of child ages may benefit from use of the GNG task, which provides relatively better precision of measurement across a broader ability range. Future studies will systematically evaluate these issues.

Longitudinal models revealed slight nonlinearity in the rate of EF ability across time; 60% of the total change in EF ability that occurred across the approximately two-year assessment period was evident in the first year. Although the nonlinear (freed loading) models provided a statistically significant improvement in fit over the linear model, globally, both models fit the data well. Moreover, overlaying the model implied trajectories for both linear and freed loading models on the observed EAP scores (Figure 3) underscore their similarity. The magnitude of change that occurs in EF ability between 3 and 5 years of age is substantial, with the average child exhibiting more than 1 standard deviation of improvement in ability over any 12 month period. These changes likely contribute to the marked improvements in children's self regulatory abilities—which are noticeable to parents and childcare providers,

alike, bolstering the idea that they are clinically meaningful—that occur during this same time period.

Similar to the conclusions drawn from longitudinal CFAs, which indicated very high correlations between EF latent factors across assessments, LGC models revealed individual differences in mean levels but not rates of change of EF ability across time. These results imply highly stable rank ordering of individual differences in EF ability across early childhood. However, rather than indicating that all children experience identical rates of improvement in EF ability across time, these results more likely suggest that, given the relatively modest amounts of latent variable score variation that are shared across tasks, there is relatively little variation left to reliably differentiate individual differences in rates of change in EF ability across time after one has taken account individual differences in levels of EF at each time.

Collectively, our results have a number of implications for practicing psychologists who are involved with the assessment of EF. First, a growing number of developmentally appropriate tasks are now available to measure EF abilities in preschool-aged children. Clinicians and researchers are no longer limited to using EF tasks that were primarily developed for adults and downwardly extended for use with preschoolers. Second, the results of this study and others like it (e.g., Wiebe et al, 2008; Hughes & Ensor, 2010) indicate that the psychometric properties of many commonly used preschool EF tasks are at least partially invariant across time. Practically speaking, there is no evidence that the EF abilities are becoming differentiated during the preschool period. Moreover, there is no reason to believe that any obtained mean level changes in EF ability during the preschool period are somehow an artifact of differential measurement. Third, all EF tasks are not “created equal”. Individual EF tasks differ in their precision of measurement. While some tasks do a good job of differentiating low levels of ability, others may do a better job of differentiating high levels of ability level (though this was less common in our specific battery). Differential measurement precision contributes to commonly observed floor and ceiling effects, which has been noted as a limitation of many commonly used EF tasks (Carlson, 2005). In the absence of tasks that have uniformly high precision of measurement across all ability levels, we advocate for the use of task batteries that involve tasks with offsetting limitations. Fourth, much of the observed variation (often 50% or more) in preschooler’s performance on any given EF task represents a combination of variation specific to that task and measurement error. Results from this study and others like it consistently imply that a child’s performance on any individual EF task is a “noisy” indicator of his/her true ability level. This, too, argues for the use of EF tasks batteries, as well as consideration of ways that children’s performance can be combined across tasks. Fifth, EF abilities are undergoing relatively rapid changes from age 3-5 years. To the extent that clinicians are implementing individualized programs designed to enhance EF abilities, simple pre-post assessments of child EF abilities represent a relatively weak standard of evidence for evaluating program effectiveness because, on average, most children will evidence improved EF abilities due simply to the passage of time. Staggered or multiple baseline designs may provide a strong approach for evaluating whether individualized programs are effective.

This study is characterized by at least three limitations. First, we prioritized the development of EF tasks that would work with young children (i.e., 3 year olds) who resided in low income households. These development efforts include extensive pilot testing of tasks with young and low-income children. As a result, the task battery does a relatively better job of measuring EF in children characterized by low to average but not high levels of ability. This may have contributed to the appearance of slowed growth in EF ability between 4-5 years versus 3-4 years of age. Second, given our focus on longitudinal measurement invariance, we presented changes in child EF ability as a function of assessment occasion. This ignores

the fact that, despite our best efforts, children's ages varied within each assessment occasion. The second-order LGC approach used here ignores this imprecision in age. Whereas most children were assessed in a narrow age window around the time of the targeted assessment, some children, whose parents were hard to reach, had their visits conducted much later than the target visit date. Although there are parameterizations of LGC models that can accommodate age variation within assessments occasions (e.g., Mehta & West, 2000), they do not conform well to our focus on testing longitudinal measurement invariance or on second-order LGC models and were thus not considered here. Subsequent studies involving this sample will provide more detailed characterization of changes in EF ability as a function of child chronological age, which may facilitate our ability to identify inter-individual differences in both level *and* rates of change in latent EF ability from 3-5 years of age (i.e., variability in both intercepts *and* slopes may be evident for growth models that are parameterized differently than those presented here). Third, this study exclusively focused on the longitudinal psychometric properties of this task battery. In the future, it will be important to both construct "short form" versions of this battery, as well as to design studies that provide 'head to head' comparisons of this task battery with other widely used EF tasks (batteries). Although psychometric adequacy is an important criterion for such comparisons, so, too, is consideration of the validity of tasks for use with under-studied populations (e.g., Autism spectrum, low birth weight, traumatic brain injury), the ability of tasks to facilitate program evaluation activities (e.g., sensitivity to change resulting from school- or individual-based intervention activities), and the prediction of longer term functioning (e.g., school readiness, psychiatric or learning disability).

The results of our first full sequence of battery development and evaluation efforts have revealed a number of limitations of the battery that currently place limits on its use in other research or clinical settings. We are currently developing a computerized version of the battery that will only require one person to administer an assessment. This change will substantially reduce data collection costs (the use of two RAs for task administration and scoring is prohibitively expensive in most settings), impose greater regularity in task administration (e.g., automate inter-stimulus interval presentation), and, through the use of touch-screen technology, facilitate the incorporation of reaction time data as an additional source of measuring individual differences in EF ability. Diamond and colleagues, among others, have demonstrated how the use of reaction time data permit the measurement of EF ability across a wide age span (Davidson et al., 2006).

A growing number of studies are identifying risk factors for preschool EF, attempting to enhance and/or remediate preschool EF, and/or using performance on preschool EF tasks to predict later developmental outcomes. A central challenge shared by many such studies is the selection of tasks that are "easy" enough for the reliable measurement of nascent abilities at age 3 (and earlier) but complex enough to delineate individual differences in rapidly developing abilities across the preschool period (and into Kindergarten). In the absence of individual tasks that evince good reliability, strong construct and predictive validity, that are developmentally scalable, and that are equally amenable for use with 3 through 5 year-old children, we believe that efforts to utilize task batteries represent a good alternative. That is, given the apparent flaws inherent in most individual EF tasks (ours included) that are used with preschool children, we advocate for the broader use of task batteries and especially the aggregation of children's performance across tasks for purposes of inference. This differs appreciably from current practice where researchers and clinicians typically select tasks perceived of as 'best' based on idiosyncratic criteria and focus on task-specific results.

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]
- Anderson, V.; Anderson, P.J.; Jacobs, R.; Spencer Smith, M. Development and assessment of executive function: From preschool to adolescence. In: Anderson, V.; Jacobs, R.; Anderson, P.J., editors. *Executive functions and the frontal lobes: A lifespan perspective*. Psychology Press; New York: 2008. p. 123-154.
- Bell MA, Wolfe CD. Changes in brain functioning from infancy to early childhood: Evidence from eeg power and coherence during working memory tasks. *Developmental Neuropsychology*. 2007; 31(1):21–38. [PubMed: 17305436]
- Bollen, KA.; Curran, P.J. *A structural equation perspective*. John Wiley and Sons, Inc; Hoboken: 2006. Latent curve models.
- Carlson SA. Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*. 2005; 28(2):595–616. [PubMed: 16144429]
- Chelune GJ, Baer RA. Developmental norms for the wisconsin card sorting task. *Journal of Clinical and Experimental Neuropsychology*. 1986; 8(3):219–228. [PubMed: 3722348]
- 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, Kirkham N, Amso D. Conditions under which young children can hold two rules in mind and inhibit a prepotent response. *Developmental Psychology*. 2002; 38:352–362. [PubMed: 12005379]
- Diamond A, Prevot MB, Callendar G, Druin DP. Prefrontal cognitive deficits in children treated early and continuously for pku. *Monographs of the Society for Research in Child Development*. 1997; Vol. 62
- Dowsett S, Livesey D. The development of inhibitory control in preschool children: Effects of “executive skills” training. *Developmental Psychobiology*. 2000; 36:161–174. [PubMed: 10689286]
- Durston S, Casey BJ. What have we learned about cognitive development from neuroimaging? *Neuropsychologia*. 2006; 44(11):2149–2157. [PubMed: 16303150]
- Durston S, Davidson MC, Tottenham N, Galvan A, Spicer J, Fossella JA, et al. A shift from diffuse to focal cortical activity with development. *Developmental Science*. 2006; 9(1):1–8. [PubMed: 16445387]
- Durston S, Thomas KM, Yang YH, Ulug AM, Zimmerman RD, Casey BJ. A neural basis for the development of inhibitory control. *Developmental Science*. 2002; 5(4):F9–F16.
- Edin F, Kilingberg T, Johansson P, McNab F, Tegner J, Compte A. Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences, USA*. 2009; 106:6810–6807.
- 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]

- Ferrer E, Balluerka N, Widaman KF. Factorial invariance and the specification of second-order latent growth models. *Methodology*. 2008; 4(1):22–36. [PubMed: 20046801]
- Fuster, JM. The prefrontal cortex. Anatomy, physiology and neuropsychology of the frontal lobe. Lippincott-Raven Press; NY: 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]
- Gerardi Caulton G. Sensitivity to spatial conflict and the development of self-regulation in children 24–36 months of age. *Developmental Science*. 2000; 3(4):397–404.
- Gerstadt C, Hong Y, Diamond A. The relationship between cognition and action: Performance of children 3 1/2 -7 years old on a stroop-like day-night test. *Cognition*. 1994; 53:129–153. [PubMed: 7805351]
- Hu, L.-t.; Bentler, PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional versus new alternatives. *Structural Equation Modeling*. 1999; 6(1):1–55.
- Hughes C, Ensor R. Individual differences in growth in executive function across the transition to school predict externalizing and internalizing behaviors and children's self-perceived academic success at age 6. *Journal of Experimental Child Psychology*. 2011; 108(3):663–676. [PubMed: 20673580]
- 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]
- Huizinga M, Dolan CV, van der Molen MW. Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia*. 2006; 44(11):2017–2036. [PubMed: 16527316]
- Jacques S, Zelazo PD. The flexible item selection task (fist): A measure of executive function in preschoolers. *Developmental Neuropsychology*. 2001; 20(3):573–591. [PubMed: 12002094]
- Jones LB, Rothbart MK, Posner MI. Development of executive attention in preschool children. *Developmental Science*. 2003; 6(5):498–504.
- Kane MJ, Engle RW. Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to stroop interference. *Journal of Experimental Psychology-General*. 2003; 132(1):47–70. [PubMed: 12656297]
- Klingberg T. Development of a superior fronto-intraparietal network for visuo-spatial working memory. *Neuropsychologia*. 2006; 44:2171–2177. [PubMed: 16405923]
- Levin HS, Culhane KA, Hartmann J, Evankovich K, Mattson AJ, Harward H, et al. Developmental-changes in performance on tests of purported frontal-lobe functioning. *Developmental Neuropsychology*. 1991; 7(3):377–395.
- Liston C, Watts R, Tottenham N, Davidson MC, Niogi S, Ulug AM, et al. Frontostriatal microstructure modulates efficient recruitment of cognitive control. *Cerebral Cortex*. 2006; 16(4):553–560. [PubMed: 16033925]
- 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]
- Luna B, Garver KE, Urban TA, Lazar NA, Sweeney JA. Maturation of cognitive processes from late childhood to adulthood. *Child Development*. 2004; 75(5):1357–1372. [PubMed: 15369519]
- Marshall PJ, Bar-Haim Y, Fox NA. Development of the eeg from 5 months to 4 years of age. *Clinical Neurophysiology*. 2002; 113(8):1199–1208. [PubMed: 12139998]
- Mehta PD, West SG. Putting the individual back into individual growth curves. *Psychological Methods*. 2000; 5(1):23–43. [PubMed: 10937321]
- Meredith W, Tisak J. Latent curve analysis. *Psychometrika*. 1990; 55(1):107–122.
- Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*. 2001; 24:167–202.
- Muthén, LK.; Muthén, BO. Mplus users guide. Fifth edition. Los Angeles, CA: 1998-2007.
- 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]

- Ram N, Grimm K. Using simple and complex growth models change: Matching theory to articulate developmental to method. *International Journal of Behavioral Development*. 2007; 31(4):303–316.
- Rebok G, Smith C, Pascualvaca D, Mirsky A, Anthony B, Kellam S. Developmental changes in attentional performance in urban children from eight to thirteen years. *Child Neuropsychology*. 1996; 2(1):1.
- Romine CB, Reynolds CR. Model of the development of frontal lobe functioning: Findings from a meta-analysis. *Applied Neuropsychology*. 2005; 12(4):190–201. [PubMed: 16422660]
- 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 S-C, Davidson MC. Memory maintenance and inhibitory control differentiate from early childhood to adolescence. *Developmental Neuropsychology*. 2010; 35(6):679–697. [PubMed: 21038160]
- Smids DP, Jacobs R, Anderson V. The object classification task for children (octc): A measure of concept generation and mental flexibility in early childhood. *Developmental Neuropsychology*. 2004; 26(1):385–401. [PubMed: 15276901]
- Stuss, D.; Knight, R., editors. *Principles of frontal lobe function*. New York: Oxford: 2002.
- Swingler MM, Willoughby MT, Calkins SD. Eeg power and coherence during preschoolers' performance of an executive function battery. *Developmental Psychobiology*. (in press).
- Thatcher RW, North DM, Biver CJ. Development of cortical connections as measured by eeg coherence and phase delay. *Human Brain Mapping*. 2008; 29:1400–1415. [PubMed: 17957703]
- Toga AW, Thompson PM, Sowell ER. Mapping brain maturation. *Trends in Neurosciences*. 2006; 29(3):148–159. [PubMed: 16472876]
- Welsh M, Pennington BF, Groisser DB. A normative-developmental study of executive function: A window on prefrontal function in children. *Developmental Neuropsychology*. 1991; 7(2):131–149.
- 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 MT, Blair CB. Test-retest reliability of a new executive function battery for sse in early childhood. *Child Neuropsychology*. 2011 DOI:10.1080/09297049.2011.554390.
- Willoughby MT, Wirth RJ, Blair CB. Contributions of modern measurement theory to measuring executive function in early childhood: An empirical demonstration. *Journal of Experimental Child Psychology*. 2011a; 108(3):414–435. [PubMed: 20553690]
- Willoughby, MT.; Wirth, RJ.; Blair, CB.; Greenberg, M.; Investigators, FLP. The measurement of executive function at age 5 years: Psychometric properties and relationship to academic achievement. 2011b. Unpublished manuscript
- Willoughby MT, Wirth RJ, Blair CB, Greenberg M, Investigators FLP. 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]
- Zelazo, PD.; Müller, U. Executive function in typical and atypical development. In: Goswami, U., editor. *Handbook of childhood cognitive development*. Blackwell; Oxford: 2002. p. 445–469.

Figure 1.
Reliability Curves for EF Tasks.

Figure 2.

Parameter estimates from freed loading second order LGC models (unstandardized coefficients). The mean and variance of the intercept, as well as all of the second-order loadings (except the loading from slope to age 3 year assessment) are fixed. Residual variances are omitted for presentation purposes.

Figure 3.
Model Implied Trajectories of Latent EF Ability Superimposed Over Observed Battery Score Estimates

\$watermark-text

\$watermark-text

\$watermark-text

Table 1

Fit statistics for item-level, longitudinal confirmatory factor analyses

Task	N	χ^2	df	p-value	CFI	RMSEA
WMS	1105	202.56	83	< 0.001	0.95	0.04
GNG	1049	129.39	98	0.018	0.99	0.02
PTP	1052	147.00	79	< 0.001	0.98	0.03
SC(A)	1117	619.72	183	< 0.001	0.93	0.05
SSS	1071	625.65	226	< 0.001	0.95	0.04
STS	1114	945.22	301	< 0.001	0.90	0.04

Note: All models estimated using the WLSMV estimator; WMS = Working Memory Span; GNG = Animal Go/No-Go; PTP = Pick the Picture; SC = Spatial Conflict; SCA = Spatial Conflict Arrows; SSS = Silly Sounds Stroop; STS = Something the Same.

\$watermark-text

\$watermark-text

\$watermark-text

Table 2

Descriptive Statistics for EAP scores for EF tasks at Each Assessment

Task	35m		48m		58m	
	M	SD	M	SD	M	SD
WMS	-0.93	0.66	-0.12	0.84	0.34	0.68
GNG	-0.43	0.95	-0.15	0.86	0.28	0.69
PTP	-	-	-0.32	0.88	0.28	0.82
SC(A) ^A	0.00	0.85	0.09	0.93	0.60	0.93
SSS	-0.51	0.80	-0.09	0.86	0.21	0.78
STS	-0.60	0.78	0.03	0.71	0.16	0.16

Note: A - No item seen at 35 months was also seen at a later assessment; WMS = Working Memory Span; GNG = Animal Go/No-Go; PTP = Pick the Picture; SC = Spatial Conflict; SCA = Spatial Conflict Arrows; SSS = Silly Sounds Stroop; STS = Something the Same.

Table 3

Nested chi-square difference tests evaluating alternative parameterization of partial measurement invariance.

Strong Invariance (constrained sub-scales)	χ^2	df	p-value	CFI	RMSEA
Model 0 (SSS, PTP, SCA, WMS, GNG, STS) ¹	252.5	116	< .001	0.89	0.03
Partially Invariant models (constrained sub-scales)					
	χ^2_{Δ}	df $_{\Delta}$	p-value	CFI	RMSEA
Model 1 (SSS, PTP, SCA, WMS, GNG)	34.51	4	< .001	0.92	0.03
Model 2 (SSS, PTP, SCA, WMS)	43.94	4	< .001	0.95	0.02
Model 3 (SSS, PTP, SCA)	19.70	4	.001	0.96	0.02
Model 4 (SSS, PTP) [*]	8.30	2	.016	0.97	0.02
Model 5 (SSS)	1.90	2	.386	0.97	0.02

Note: N = 1121; chi square test statistic and probability value for model 0, represents a test of a longitudinal strong measurement invariance against the observed data; in contrast, chi square test statistics for models 1-5 represent scaled nested tests of the relative fit of models in which fewer EF tasks are assumed to be invariant across assessments (e.g., model 2 represents the improvement in fit, relative to model 1, that results by allowing the factor loadings and task intercepts of the GNG task to vary across assessments);

^{*} final selected model; CFI = Comparative Fit Index; RMSEA = Root Mean Squared Error of Approximation; WMS = Working Memory Span; GNG = Animal Go/No-Go; PTP = Pick the Picture; SCA = Spatial Conflict Arrows; SSS = Silly Sounds Stroop; STS = Something the Same.