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

Michael T. Willoughby
University of North Carolina at Chapel Hill

R. J. Wirth
Vector Psychometric Group

Clancy B. Blair
New York University

Family Life Project Investigators

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 = 1,292) of families who were oversampled from low-income and African American families at the birth of a new child—at assessments conducted when the child was 3, 4, and 5 years old. All 6 individual EF tasks exhibited strong measurement invariance over time. The EF battery, which was derived from the 6 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- and 5-year assessments occurred between the Year 3 and Year 4 assessments. Results are discussed with respect to the importance of establishing scalable measures of EF ability prior to investigating experiences that predict 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, Arns, Anderson, & Diamond, 2006; Diamond, Kirkham, & Arns, 2002; Garon, Bryson, & Smith, 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 have shown that EF is meaningfully related to a number of aspects of child development including social–emotional competence (Carlson, Mandell, & Williams, 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, Müller, Frye, & Marcovitch, 2003), measures of EF abilities in young children, 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, presented a variety of tasks in a uniform format, was
easily administered by lay staff, elicited individual differences in ability level, and 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- to 5-year-olds, as well as at individual time points, ages 3, 4, and 5 years, in the large, representative sample of children analyzed longitudinally here (Willoughby & Blair, 2011; Willoughby, Wirth, & Blair, 2011; Willoughby, Wirth, Blair, Greenberg, & Family Life Project investigators, 2010; 2011). 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 (Espe, Sheffield, Wiebe, Clark, & Moehr, 2011; Hughes, Ensor, Wilson, & Graham, 2010; Shing, Lindenberger, Diamond, Li, & Davidson, 2010; Wiebe, Espe, & Charak, 2008). Second, whereas the short-term (approximately 2–3 weeks) test–retest reliability of individual tasks in our battery was poor to modest, rs ≈ .60, the test–retest reliability of the battery was excellent when examined with a latent variable model, φ = .95. This result underscores an underappreciated 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 to .50). Aggregating performance across tasks improves the precision of measurement of EF ability and is preferable to 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 3-year-olds) who resided in low-income homes. The primary objectives of this current study were (a) to extend our previous psychometric work by testing for longitudinal measurement invariance of our individual tasks, as well as the task battery scores, and (b) to describe the developmental trajectory of EF abilities as measured by our task battery from ages 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 (Durston & Casey, 2006; Durston 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, Thompson, & Sowell, 2006), precise longitudinal measurement of EF 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 & Buer, 1986; Davidson et al., 2006; Huizinga, Dolan, & van der Molen, 2006; Levin et al., 1991; Luna, Garver, Urban, Lazar, & Sweeney, 2004; Rebek et al., 1996; Welsh, Pennington, & Groisser, 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 ages 3–5 years. Early childhood is understood to be a period of pronounced developmental improvements in EF abilities (Anderson, Anderson, Jacobs, & Spencer Smith, 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, Hong, & Diamond, 1994; Jones, Rothbart, & Posner, 2003; Luciana & Nelson, 1998; Smidts, Jacobs, & Anderson, 2004). Secondary evidence comes from studies that have employed electroencephalograph (EEG) methods. These studies have documented changes in frontal-based circuitry during the early childhood period (Marshall, Hart-Haim, & Fox, 2002; Thatcher, North, & Biver, 2008), as well as evidence that individual differences in EEG power or coherence during the completion of EF tasks are associated with enhanced EF performance (Bell & Wolfe, 2007; Swingler, Willoughby, & Calkins, in press).

To the best of our knowledge, Hughes, Ensor, and colleagues (Hughes & Ensor, 2011; Hughes et al., 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, Prevor, Callendar, & Druin, 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 interindividual differences in intraindividual change), the ability to use growth parameters (e.g., intercepts, slopes) as both independent and dependent variables. One of the strengths of Hughes and Ensor’s (2011) 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 assessments conducted at ages 4 and 6 years. 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 children’s 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 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 interindividual differences in intraindividual change of latent EF ability between 3 and 5 years of age (Bollen & Curran, 2006; Ferrer, Balluerka, & Widaman, 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 to 5 years.

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 1,292 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 North Carolina were oversampled (African American families were not oversampled in Pennsylvania 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 Project Investigators, in press).

Of those families interested and eligible and selected to participate in the study, 1,292 families completed a home visit at 2 months of child age, at which point they were formally enrolled in the study. A total of $N = 1,123$ (87%), $N = 1,066$ (83%), and $N = 1,099$ (85%) families participated in the age 3-, 4-, and 5-year assessments, respectively. While most families ($N = 1,017$; 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 ($N = 1,168$) 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 Pennsylvania, $p = .67$), or being recruited from the low-income stratum (76% vs. 78% poor, $p = .62$).

Procedures

Families participated in home visits when study children were 3 (two visits), 4 (one visit) and 5 (one 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 2 hr to complete, of which children spent approximately 30–45 min (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 flip-book format (pages measured 8 in. $\times$ 14 in.), 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 as many as 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 or her; 2 = average quality, indicated by few distractions and confidence that the child understands what is being asked of him or 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 or her performance. Brief descriptions of each task are provided in the following sections. 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 in the pretest phase that the child knows both colors and animals, the examiner asks the child to name the animal and then to name the color. The examiner then turns the page, which shows only the outline of the house from the previous page. The examiner then asks the child which animal was in or 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 1 one-house trial, 2 two-house trials, 3 three-house trials, and 2 four-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., due to the results of pilot testing and test burden, the four-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 two-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 two-picture, three-picture, four-picture, and six-picture sets (the first item in each set was not scorable 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 Gerhardt, 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 are 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 eight 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 turns flip-book pages that depict either a car or boat. The child is instructed to touch the car on his or her response card when the flip-book page shows a car and to touch the boat on the response card when the page shows a boat. Across the first eight trials, cars and boat are 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 right hand 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 contralaterally, 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 contralaterally require inhibitory control from the previously established prepotent response in order to be answered correctly (spatial location is no longer informative). Responses (correct, incorrect) to contralaterally 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 earlier, 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 eight 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 contralaterally, 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 contralaterally require inhibitory control from the previously established prepotent response in order to be answered correctly (spatial location is no longer informative). Responses (correct, incorrect) to contralaterally presented items were used for purposes of scoring. The SCA was administered at the 4- and 5-year assessments.

**Animal Go/No-Go (GGN; inhibitory control).** This is a standard go/no-go task (e.g., Durston et al., 2002) presented in a flip-book 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 2 s, with each page depicting a line drawing of one of seven possible animals. The task presents varying numbers of go trials prior to each no-go trial, including, in standard order, one-go, three-go, three-go, five-go, one-go, one-go, and three-go trials. Responses (correct, incorrect) to no-go trials were used for purposes of scoring. The GGN 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 are 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 or 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, CFA’s 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, CFA’s 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 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 freed-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 that 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.

For the IRT models, we used Bock–Aitkin expectation-maximization estimation as implemented in IRTPRO (Cai, du Toit, & Thissen, in press). IRT models included either two-parameter logistic models (2PLMs) for those tasks with dichotomous items (i.e., SC, SCA, SSS, GGN, STS) or graded responses models for those tasks with polytomous item response formats (i.e., PTP, WMS). Consistent with earlier reports, two of the tasks involved bifactor 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 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) ≥ .95 *and* root-mean-square errors of approximation (RMSEA) ≤ .05 were considered to exhibit good fit, while those models with CFI ≥ .95 or RMSEA ≤ .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 oversampling) but battery-level analyses (CFA and LGC models) did.

**Results**

**Participation in EF Assessments**

As we have described in companion articles, 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 at 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, or physical impairments that prohibited their participation (Willoughby et al., 2010).

In total, 973 (75% of the total sample), 1,009 (78%), and 1,038 (80%) children participated in EF testing at 3-, 4-, and 5-year home visits, respectively. A total of 865 (67% of the total sample) participated in EF testing at all three assessments, 169 (13%) at two of the three assessments, 87 (7%) at one of the three assessments, and 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 nonparticipation in these assessments; \( N = 171 \)) did not differ from children who did participate in at least one EF assessment (\( N = 1,121 \)) 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 Pennsylvania, \( p = .26 \)), or being recruited from the low-income stratum (77% vs. 78% poor, \( p = .75 \)).

The 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 1,009 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 1,038 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 bifactor 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 task 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 Wirth, & Blair, 2011). 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 \( N_s = 929 –1,045 \)). 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., Curran et al., 2008; Hussong et al, 2007). 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.

**Table 1**

<table>
<thead>
<tr>
<th>Task</th>
<th>( N )</th>
<th>( \chi^2 )</th>
<th>( df )</th>
<th>( p ) value</th>
<th>CFI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMS</td>
<td>1,105</td>
<td>202.56</td>
<td>83</td>
<td>(&lt;0.001)</td>
<td>0.95</td>
<td>0.04</td>
</tr>
<tr>
<td>GNG</td>
<td>1,049</td>
<td>129.39</td>
<td>98</td>
<td>0.018</td>
<td>0.99</td>
<td>0.02</td>
</tr>
<tr>
<td>PTP</td>
<td>1,052</td>
<td>147.00</td>
<td>79</td>
<td>(&lt;0.001)</td>
<td>0.98</td>
<td>0.03</td>
</tr>
<tr>
<td>SC(A)</td>
<td>1,117</td>
<td>619.72</td>
<td>183</td>
<td>(&lt;0.001)</td>
<td>0.93</td>
<td>0.05</td>
</tr>
<tr>
<td>SSS</td>
<td>1,071</td>
<td>625.65</td>
<td>226</td>
<td>(&lt;0.001)</td>
<td>0.95</td>
<td>0.04</td>
</tr>
<tr>
<td>STS</td>
<td>1,114</td>
<td>945.22</td>
<td>301</td>
<td>(&lt;0.001)</td>
<td>0.90</td>
<td>0.04</td>
</tr>
</tbody>
</table>

**Note.** All models estimated using the weighted least squares with mean and variance adjustment 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’s the Same game.

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-year to 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 (vs. 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

### Table 2

Descriptive Statistics for EAP Scores for EF Tasks at Each Assessment

<table>
<thead>
<tr>
<th>Task</th>
<th>35 months</th>
<th>48 months</th>
<th>58 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>WMS</td>
<td>-0.93</td>
<td>0.66</td>
<td>-0.12</td>
</tr>
<tr>
<td>GNG</td>
<td>-0.43</td>
<td>0.95</td>
<td>-0.15</td>
</tr>
<tr>
<td>PTP</td>
<td>—</td>
<td>—</td>
<td>-0.32</td>
</tr>
<tr>
<td>SCA(A)*</td>
<td>0.00</td>
<td>0.85</td>
<td>0.09</td>
</tr>
<tr>
<td>SSS</td>
<td>-0.51</td>
<td>0.80</td>
<td>-0.09</td>
</tr>
<tr>
<td>STS</td>
<td>-0.60</td>
<td>0.78</td>
<td>0.03</td>
</tr>
</tbody>
</table>

* No item seen at 35 months was also seen at a later assessment.

Note. 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’s the Same game.

Figure 1. Reliability curves for executive function tasks. GNG = animal go/no-go; SCA = spatial conflict arrow; SC = spatial conflict; PTP = pick the picture; STS = Something’s the Same game; SSS = Silly Sounds Stroop; WMS = working memory span.
(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, SCA, 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, $\chi^2(116) = 252.5, p < .001, \text{CFI} = 0.89, \text{RMSEA} = 0.03$. 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, $\chi^2(102) = 145.4, p = .003, \text{CFI} = 0.97, \text{RMSEA} = 0.02$.

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 to .60 (task $R^2 = .04$–.36), consistent with only modest intercorrelations 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: $\varphi_{3 \& 4 \text{ years}} = .88, \varphi_{3 \& 5 \text{ years}} = .86, \varphi_{4 \& 5 \text{ years}} = .91, ps < .001$, 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 that 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, $\chi^2(105) = 163.2, p = .0002, \text{CFI} = 0.95, \text{RMSEA} = 0.02$. The linear slope term was statistically significant ($b_{\varphi} = 0.12, p < .001$), indicating model-implied improvements of approximately one tenth 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 interindividual differences in intraindividual change in true EF ability level. The inability to detect individual differences in rates of change in performance across time explained the nonsignificant correlation between growth parameters ($\varphi = .49, p = .43$).

The second-order freed-loading LGC model also fit the data well, $\chi^2(104) = 151.6, p = .002, \text{CFI} = 0.96, \text{RMSEA} = 0.02$. The estimated factor loading ($\lambda$) for the 3-year assessment was $-18.2, p < .001$—compared with a fixed value of $-12$ in the linear LGC. Using an equivalent parameterization, the estimated factor loading for the 5-year assessment resulted in $7.3, p < .001$—compared with 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, $\Delta \chi^2(1) = 9.1, p = .003$. 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- and 5-year assessments occurred between the 3-year 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 cali-

### Table 3

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$p$</th>
<th>CFI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly invariant (constrained subscales)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 0 (SSS, PTP, SCA, WMS, GNG, STS)</td>
<td>252.5</td>
<td>116</td>
<td>&lt;.001</td>
<td>0.89</td>
<td>0.03</td>
</tr>
<tr>
<td>Partially invariant (constrained subscales)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1 (SSS, PTP, SCA, WMS, GNG)</td>
<td>34.51</td>
<td>4</td>
<td>&lt;.001</td>
<td>0.92</td>
<td>0.03</td>
</tr>
<tr>
<td>Model 2 (SSS, PTP, SCA, WMS)</td>
<td>43.94</td>
<td>4</td>
<td>&lt;.001</td>
<td>0.95</td>
<td>0.02</td>
</tr>
<tr>
<td>Model 3 (SSS, PTP, SCA)</td>
<td>19.70</td>
<td>4</td>
<td>&lt;.001</td>
<td>0.96</td>
<td>0.02</td>
</tr>
<tr>
<td>Model 4 (SSS, PTP)</td>
<td>8.30</td>
<td>2</td>
<td>.016</td>
<td>0.97</td>
<td>0.02</td>
</tr>
<tr>
<td>Model 5 (SSS)</td>
<td>1.90</td>
<td>2</td>
<td>.386</td>
<td>0.97</td>
<td>0.02</td>
</tr>
</tbody>
</table>

**Note.** $N = 1,121$; 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 nested tests of the related fit of models in which fewer executive function 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 animal go/no-go task to vary across assessments); * final selected model; CFI = comparative fit index; RMSEA = root mean squared error of approximation; SSS = Silly Sounds Stroop; PTP = pick the picture; SCA = spatial conflict arrows; WMS = working memory span; STS = Something's the Same game.
In a free-loading 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% and 38% of the observed variation in task (EAP) scores was explained by latent ability factors, between 89% and 91% of the observed variation in the battery scores was 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 and 4 years relative to 4 and 5 years of age. Although there were stable interindividual 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 (2011) 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 and 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 ≥.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- through 5-year-olds or children older than 5 years). Reliability curves are directly related to item

![Figure 3. Model implied trajectories of latent executive function ability superimposed over observed battery score estimates. EF = executive functioning; EAPs = expected a posteriori.](image-url)
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 later, 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 CFAs, in which children’s performance on individual tasks at each assessment served as indicators of latent EF ability at each assessment, demonstrated that 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 or PTP tasks would appear prudent, given evidence that these are equally good indicators of EF ability from 3 to 5 and from 4 to 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. The issues will be systematically evaluated in future studies.

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 2-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 child care 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., Hughes & Ensor, 2011; Wiebe et al., 2008) 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 preschoolers’ 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 or 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 to 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 pro-
vide 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 and 5 years versus between 3 and 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 interindividual differences in both level and rates of change in latent EF ability from 3 to 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 both to construct “short-form” versions of this battery and 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 understudied 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 now are developing a computerized version of this 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. Davidson et al. (2006), among others, have demonstrated how the use of reaction time data permit the measurement of EF ability across a wide age span.

A growing number of studies are identifying risk factors for preschool EF, attempting to enhance or remediate preschool EF, or using performance on preschool EF tasks to predict later development 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.

References
Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J.,...


Received June 18, 2011
Revision received July 28, 2011
Accepted August 1, 2011