The Measurement of Executive Function at Age 3 Years: Psychometric Properties and Criterion Validity of a New Battery of Tasks

Michael T. Willoughby  
University of North Carolina at Chapel Hill

Clancy B. Blair  
New York University

R. J. Wirth  
University of North Carolina at Chapel Hill

Mark Greenberg  
Pennsylvania State University

The Family Life Project Investigators

In this study, the authors examined the psychometric properties and criterion validity of a newly developed battery of tasks that were designed to assess executive function (EF) abilities in early childhood. The battery was included in the 36-month assessment of the Family Life Project (FLP), a prospective longitudinal study of 1,292 children oversampled from low-income and African American families. Ninety-one percent of children were able to complete 1 or more of the tasks. Psychometric analyses were used to test the dimensionality of each task, evaluate the item and task properties, test the dimensionality of the task battery, and evaluate the criterion validity of the battery with multi-informant measures of attention-deficit/hyperactivity disorder (ADHD) symptomatology and child performance on two subtests of the Wechsler Preschool and Primary Scale of Intelligence. Results indicated that the tasks were successful in measuring interindividual differences in child EF ability, that task scores were most informative about ability level for children in the low to moderate range of ability, that children’s performance across the entire battery was adequately summarized by a single factor, and that individual differences on the EF battery were related to ADHD symptomatology and intelligence in expected ways. Results are discussed with respect to the importance of developing psychometrically sound, scalable instruments that facilitate the measurement of interindividual differences in intrindividual change of EF across the early childhood period.

Keywords: executive function, early childhood, psychometrics

Supplemental materials: http://dx.doi.org/10.1037/a0018708.supp

Executive function (EF) is an umbrella term that refers to a wide range of cognitive abilities—including attentional control, cognitive flexibility, goal setting, and specific aspects of information processing including fluency and processing speed—that are involved in the control and coordination of information in the service of goal-directed actions (Anderson, 2002; Fuster, 1997; Miller & Cohen, 2001). As such, EF can be defined as a supervisory system that is important for planning, reasoning ability, and the integration of thought and action (Shallice & Burgess, 1996). At a more fine-grained level, however, EF, as studied in the cognitive development literature, has come to refer to specific interrelated information processing abilities that enable the resolution of conflicting information, namely, working memory (WM), defined as the holding in mind and updating of information while performing some operation on it; inhibitory control, defined as the inhibition of prepotent or automatized responding when engaged in task completion; and attention shifting, defined as the ability to shift cognitive set among distinct but related dimensions or aspects of a given task (Davidson, Amso, Anderson, & Diamond, 2006; Miyake, Friedman, Emerson, Witzki, & Howarter, 2000; Zelazo & Muller, 2002). Focusing on these more narrowly defined abilities is particularly apropos when studying EF in early childhood, as many of the more complex aspects of EF (e.g., abstract thought; goal setting) have an extended developmental course and are not easily measured in very young children (Garon, Bryson, & Smith, 2008).
One of the reasons for widespread interest in EF is its relation to brain development. EFs are closely associated with prefrontal cortex (Miller & Cohen, 2001; Stuss & Knight, 2002), an area of the brain with a protracted developmental timetable extending into early adulthood (Toga, Thompson, & Sowell, 2006). In keeping with knowledge of prefrontal cortex development, cross-sectional studies have shown that performance on relatively low demand EF tasks asymptotes early, with mature levels of performance reached at approximately mid-childhood (Luciana & Nelson, 1998). Performance on tasks with more complex processing demands, however, continues to increase throughout adolescence, reaching mature levels only in young adulthood (Davidson et al., 2006; Luciana, Conklin, Hooper, & Yarger, 2005).

In addition to the association with brain development, the construct and criterion validity of EF tasks in early childhood have been established through relations of EF with several aspects of child development, including temperament (Rothbart, Ellis, Rueda, & Posner, 2003; Gerardi-Caulton, 2000), social–emotional development (Carlson, Mandell, & Williams, 2004; Hughes & Ensor, 2007), academic ability (Blair & Razza, 2007; Bull & Scerif, 2001; Diamond, Barnett, Thomas, & Munro, 2007), and general intelligence and verbal ability (Carlson, Moses, & Breton, 2002). Nonetheless, there is a need for more careful psychometric research on EF measures in order to better understand both individual differences and developmental course (Blair, Zelazo, & Greenberg, 2005).

Carlson (2006) summarized numerous tasks that have been developed to measure EF ability in early childhood. Children’s performance on many of these tasks has provided an empirical basis for documenting improvement in EF ability during the early childhood period. Nonetheless, collectively, the extant measures of EF in early childhood suffer from at least four limitations. First, many tasks were developed and/or scored to demonstrate skill acquisition in dichotomous terms (children of varying ages either pass or fail tasks). Conceptualizing skill acquisition in dichotomous terms masks individual differences in EF ability. Second, most tasks were developed for use in laboratory settings. As such, they have often been administered by highly trained staff (graduate students), often lack clearly written training protocols, and involve idiosyncratic test materials. Together, this limits their wide-scale use by trained lay interviewers in the context of large-scale studies. Third, most tasks were developed with small convenience samples of children whose parents were motivated to participate in university research. Hence, the generalizability of extant tasks for use with economically disadvantaged or disordered populations is unknown. Fourth, most extant EF tasks for use in early childhood have not undergone formal psychometric evaluations. This vitiated our ability to test theoretical questions related to the developmental course, as well as the criterion and predictive validity of EF abilities. It would be inaccurate to imply that all EF tasks that are available for use in early childhood suffer from these limitations. For example, the shape school task includes scoring that facilitates understanding of individual differences (not simply pass–fail decision making) and has undergone psychometric evaluations in a reasonably large sample (Espy, 1997; Espy, Bull, Martin, and Stroup, 2006). However, to the best of our knowledge, the task has always been administered by highly trained graduate students in the context of a lab-based assessment. Moreover, this task is an exception rather than the rule.

We have been working to develop a battery of EF tasks that overcome these limitations. Specifically, we have created and/or adapted five tasks that putatively measure three dimensions of EF in early childhood: WM, inhibitory control, and attention shifting. The first task is a spanlike measure of WM (Engle, 2002) in which children are required to name two objects, to hold them both briefly in mind, and to name one and inhibit the other while both are hidden from view. The second task is an IS task (adapted from Jacques and Zelazo 2001) in which the child’s attention is initially focused on one way in which two of three items are similar before he or she is asked to identify a second way in which two of the three items are similar. The third task is a spatial conflict (SC) task (adapted from Gerardi-Caulton, 2000) in which responding is established on trials in which stimulus and response are compatible and then inhibited on trials in which stimulus and response are incompatible. The fourth task is a Stroop-like inhibitory control task in which a prepotent response must be inhibited in favor of a rule based subdominant response. The fifth task is a go/no-go (GNG) task in which a response must be inhibited to only one infrequently occurring item in a specified set of stimuli.

The set of EF tasks shared a number of features. Each task was presented in a common flipbook format that included highly scripted instructions that were read to children by a research assistant (RA). Each task was administered by two RAs, one who was responsible for administering tasks to the child and the other who was responsible for recording child responses (but not evaluating the accuracy of responses). By disassembling administration and response recording roles, and not requiring either RA to evaluate the accuracy of child responses (accuracy was evaluated with computerized scoring), we minimized the demand on RAs, making the tasks more amenable to administration by staff who have little specialized training or expertise in task content. To be clear, we are not suggesting that it is desirable to have EF tasks administered by individuals who lack specialized training or expertise in task content. However, there are numerous circumstances (e.g., large-scale research projects; school wide assessments) in which the measurement of EF by specialists is impractical or unfeasible. In those situations, it is important to have tasks that can be used by lay staff after minimal training. Iterative pilot testing of tasks with nearly 300 primarily low-income families helped us identify items and tasks that were too difficult for young children, as well as revealed text that required edits in order to facilitate comprehension. Following pilot testing, the battery was included at the 36-month assessment of an ongoing prospective longitudinal study (N = 1,292) of families who were recruited from central Pennsylvania and eastern North Carolina, with oversampling for low-income and, in North Carolina, African American families.

The primary goal in this study is to evaluate the psychometric properties of each of the EF tasks that were administered at the 36-month assessment. The secondary and tertiary goals in this study were to evaluate the dimensionality of the battery of tasks and to evaluate the criterion validity of the task battery. With respect to the dimensionality of EF, research conducted with adolescent and young adult samples has indicated that EF is best conceptualized as multidimensional, with distinct factors for WM, inhibitory control, and attention shifting abilities, despite the fact that most EF tasks make integrative demands on these sets of abilities (Friedman et al., 2006; Miyake et al., 2000). Although
there are theoretical arguments to support the conceptualization of
EF as a multidimensional construct in early childhood (Garon et
al., 2008 for review), the available empirical evidence indicates
that EF is best characterized as a unitary factor (Carlson, Mandell,
& Williams, 2004; Hughes & Ensor, 2007; Wiebe, Espy, &
Charak, 2008). On the basis of these results, we hypothesized that
performance on the set of tasks in the battery would be adequately
characterized by a single-factor model.

Criterion validity was evaluated by correlating individual dif-
fferences in performance on the EF task battery with multi-
informant measures of attention-deficit/hyperactivity disorder
(ADHD) symptomatology and child performance on two subtests
of the Wechsler Preschool and Primary Scale of Intelligence
(WPPSI–III), which at age 3 years can also be used to obtain a full
scale estimate of IQ (Sattler, 2001). EF figures prominently in
modern theoretical accounts of ADHD (Barkley, 1997; Nigg,
Willcutt, Doyle & Sonuga-Barke, 2005). Moreover, results from
two recent meta-analyses indicated moderate to large differences
between ADHD youths and non-ADHD youths on multiple indices
of EF, including response inhibition, vigilance, WM, and planning
(Martinussen, Hayden, Hogg-Johnson, & Tannock, 2005; Willcutt,
Doyle, Nigg, Faraone, & Pennington, 2005). As such, we hypoth-
esized that children’s performance on the EF task battery would be
negatively correlated with ADHD symptomatology. Moreover,
although EF is understood to be distinct from general intelligence
(Blair, 2006), WM, a specific component of EF, is highly corre-
lated with general intelligence in latent variable analyses, with a
median correlation of .72 (Kane, Hambrick, & Conway, 2005).
Although the strong association between EF and IQ has been
established that children’s performance on the EF task battery would be
positively correlated with IQ as well.

Method

Participants

The Family Life Project (FLP) was designed to study young
children and their families who lived in two of the four major
gеographical 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 used 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 over-
sampled (African American families were not oversampled in
Pennsylvania because the target communities were at least 95% non–African American).

At both sites, recruitment occurred 7 days per week over the
12-month recruitment period spanning September 15, 2003,
through September 14, 2004, with a standardized script and
screening protocol. The coverage rate was over 90% for all births
that occurred to women in these counties in that 1-year period. In
Pennsylvania, families were recruited in person from three hos-
pitals. These three hospitals represented a weighted probability
sample (hospitals were sampled proportional to size within county)
of seven total hospitals that delivered babies in the three target
Pennsylvania counties. Pennsylvania hospitals were sampled be-
cause the number of babies born in all seven target hospitals far
exceeded the number needed for purposes of the design. In North
Carolina, families were recruited in person and by phone. In-
person recruitment occurred in all three of the hospitals that
delivered babies in the target counties. Phone recruitment occurred
for families who resided in target counties but delivered in non-
target county hospitals. These families were located through sys-
tematic searches of the birth records located in the county court-
houses of nearby counties.

FLP recruiters identified 5,471 (59% North Carolina, 41%
Pennsylvania) women who gave birth to a child in the 12-month
period. A total of 1,515 (28%) of all identified families were
determined to be ineligible for participation for three primary
reasons: not speaking English as the primary language in the home,
residence in a nontarget county, and intent to move within 3 years.
Of the 2,691 eligible families who agreed to the randomization
process, 1,571 (58%) families were selected to participate with the
sampling fractions that were continually updated from our data
center. Of those families selected to participate in the study, 1,292
(82%) families completed a home visit at 2 months of child age, at
which point they were formally enrolled in the study. The current
study focused on children’s performance on a newly developed
battery of EF tasks that were administered at the 36-month home
visit. Families and children who participated in the 36-month visit
(n = 1,123) did not differ from those who did not participate in this
visit (n = 169) with respect to state of residence (40% vs. 38%
residing in Pennsylvania, p = .63), race of the child (43% vs. 40%
African American, p = .42), or being recruited in the low-income
stratum (77% vs. 79% poor, p = .70). However, children of the
families who participated in the 36-month visit were more likely to
be female than were children of the families who did not partici-
pate in the 36-month visit (50% vs. 41%, p = .02).

Procedures

Families participated in two home visits when children were
approximately 36 months of age. At the first visit, children com-
pleted standardized testing (including the WPPSI–III task de-
scribed below). At the second visit, children were administered
five newly developed tasks that were designed to measure their EF
ability. Children were seated across from the experimenter at a
convenient location in the home. One RA was responsible for
administering EF tasks (in a fixed order) to children, including
keeping them engaged and making decisions about how frequently
to take breaks. A second RA was responsible for recording chil-
dren’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. The EF tasks were
administered at the conclusion of an assessment session in which
children also completed a series of tasks with the mother that
included a picture-book reading task, an empathy task, and a
puzzle task. Cumulatively, these tasks took about 1 hour to com-
plete. Parents completed interviews and questionnaires at both
visits. Home visitors independently completed impressions of child behavior at the conclusion of each visit.

**Measures**

**EF.** Each of the EF tasks was presented in an open spiral bound flipbook format that allows the examiner to easily turn pages that present stimuli on one page and scripted instructions for administration on the other. Pages measure 8 in. × 14 in. For each of the tasks, 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.

The WM task is a spanlike measure that is based on 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 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 that only shows the outline of the house from the previous page. The examiner then asks the child which animal was 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, and two 3-house trials.

The SC task is a Simon task similar to that used by Gerardi-Caulton (2000), which is intended to assess inhibitory control. 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 pages that depict either a car or boat. The child is instructed to touch the car on the response card when the 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 boats 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 flipbook page (above the car on the response card) and boats always appearing on the right side of the flipbook page (above the boat on the response card). These items build a prepotency to touch the response card on the basis of the location (images on the left or right of each flipbook page always result in the child touching the left or right side of the response card). 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 flipbook page (above the boat on the response card) and boats appearing on the left side of the flipbook page (above the car on the response card). Items presented contra-laterally require inhibitory control from the previously established prepotent response in order to be answered correctly (spatial location is no longer informative).

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

The silly sounds task is a Stroop-like task similar to that developed by Gerstadt, Hong, and Diamond (1994) that is intended to assess inhibitory control of a prepotent response. In this task, children are instructed to make the sound that a dog makes in response to a line drawing of a cat and to make the sound that a cat makes in response to a line drawing of a dog. Following a pretest phase, children are presented with 18 trials in which a line drawing of a dog is presented beside a line drawing of a cat.

The animal GNG is a standard GNG (e.g., Durston et al., 2002) task that is intended to assess inhibitory motor control. 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, five-go, five-go, one-go, one-go, and three-go trials.

**WPPSI–III** (Wechsler, 2002). Children completed the Vocabulary and Block Design subscales of the WPPSI–III in order to provide an estimate of intellectual functioning at age 36 months (Sattler, 2001).

**ADHD symptom rating checklist** (DuPaul et al., 1998). The ADHD checklist includes the 18 Diagnostic and Statistical Manual of Mental Disorders (4th edition; American Psychiatric Association, 2000) symptoms for ADHD, each rated on a 4-point scale (0 = not at all, 1 = just a little, 2 = pretty much, 3 = very much). The internal consistency of the set of 18 items was strong (α = .93). Following convention for the use of this instrument and others like it (see, for example, Pelham et al., 1992), items that were rated by parents as either pretty much or very much describing the behavior of their child were considered endorsed symptoms. A summary variable representing the total number of endorsed ADHD symptoms was used in criterion validity analyses.

**Strengths and Difficulties Questionnaire** (SDQ; R. Goodman, 1997). The SDQ is a norm-referenced, 25-item behavior rating scale that yields five 5-item scales, including Prosocial Behavior, Peer Problems, Emotional Problems, Hyperactivity, and Conduct Problems. The reliability and the validity of the SDQ are well established (R. Goodman, 2001; A. Goodman & Goodman, 2009). The child’s primary daycare provider (when applicable) and two separate RAs, who participated in home visits, completed the Hyperactivity subscale of the SDQ. The internal consistency of the Hyperactivity subscale for childcare providers was good (α = .74). Four ratings from RAs (independent ratings from each of two RAs...
at two separate home visits) were summarized as a single mean score for use in the current study (rs = .57 & -.67; α = .86). Daycare provider and RA ratings of Hyperactivity supplemented parent reports of ADHD in criterion validity analyses.

Analytic Strategy

Analyses proceeded in four phases. First, confirmatory factor analyses (CFAs) were used to evaluate the dimensionality of each EF task. Each task was developed to be unidimensional. However, when the fit of unidimensional models was poor, bifactor models were considered. Bifactor models introduce method factors that take into account residual correlations that exist between items, even after accounting for their covariation due to a shared general factor. Second, item response theory (IRT) models were applied to each task for purposes of item parameter estimation, task scoring (i.e., computation of expected a posteriori [EAP] estimates), and evaluating test information functions. Third, CFA models were used to evaluate the dimensionality of the entire EF battery, with EAPs from IRT models as the indicators. Fourth, CFAs were used to evaluate the criterion validity of the EF task battery by relating a latent variable representing child performance on the EF task battery to latent variables representing child performance on IQ subtests, as well as multi-informant ratings of behaviors characteristic of ADHD. In addition, a logistic regression model was estimated to ascertain whether the EF latent variable was predictive of the presence of six or more inattentive and/or hyperactive–impulsive parent-rated symptoms. CFA models were estimated with Mplus (Version 5; Muthén & Muthén, 2006). IRT models were estimated with methods outlined by Gibbons and Hedeker (1992, see also, Gibbons et al. 2007) as implemented in IRTPro (Beta version; Cai, du Toit, & Thissen, in press). The logistic regression model was estimated with PROCs LOGISTIC and SURVEYLOGISTIC in SAS (version 9.1). Unless otherwise noted, all analyses took into account the complex sampling design (stratification, oversampling of low-income and, in North Carolina, African American families).

Results

Sample Description and Rates of EF Task Completion

Descriptive characteristics of the families and children who participated in the 36-month visit are provided in Table 1. Of the 1,123 children and families who participated in the 36-month visit, 54 (5%) children did not have the opportunity to complete the EF battery. This was due primarily to families moving out of the geographic area and having interviews conducted by phone (no opportunity for child testing), but also included instances in which only one of the two planned 36-month visits could not be completed. With three exceptions, children who were not given an opportunity to complete EF tasks were indistinguishable from children who completed one or more tasks with respect to demographic factors (see Table 1). Specifically, children who were not given an opportunity to complete tasks were more likely to reside in North Carolina (83% vs. 59%, p = .0004), to have primary caregivers who were married (70% vs. 58%, p = .06), and to be older (M = 38.6 vs. M = 37.0 months, p = .06) than children who completed one or more tasks.

Of the remaining 1,069 children, 94 (8% of all children; 9% of those given opportunity) children were unable to complete any of the EF tasks, whereas 975 (87% of all children; 91% of those given opportunity) children completed one or more tasks. Compared with children who completed one or more EF tasks, child who were unable to completed any tasks were more likely to reside in Pennsylvania (51% vs. 41%, p = .05); to be male (77% vs. 47%, p < .0001); to be younger (M = 36.5 vs. M = 37 months, p = .0003); to have a primary caregiver who was unmarried (57% vs. 42%, p = .005); to reside in lower income households (M = 1.3 vs. M = 1.9 income/needs ratio, p < .0001); to have been rated as more hyperactive by their primary caregiver (M = 5.7 vs. M = 4.0, p = .001), by their childcare provider (M = 1.1 vs. M = 0.8, p = .001), and by RAs who participated in home visits (M = 1.3 vs. M = 0.6, p < .0001); and to have scored lower on the Receptive Vocabulary (M = 83.4 vs. M = 100.2, p < .0001) and Block Design (M = 76.8 vs. M = 91.6, p < .0001) subtests of the WPPSI–III, with a corresponding difference on the estimated full scale IQ score (M = 75.9 vs. M = 95.1, p < .0001). The WPPSI–III scores, in particular, indicate that children who were unable to complete any EF tasks were developmentally delayed. Descriptive statistics are summarized in Table 1.

Among the 975 children who completed at least one EF task, children completed an average of 3.6 (median = 4) of 5 possible tasks. Specifically, 30% completed all 5 tasks, 27% completed 4 tasks, 24% completed 3 tasks, 11% completed 2 tasks, and 8% completed only a single task. Rates of individual task completion varied substantially: 94% completion of SC, 87% completion of IS, 83% completion of WM span, 51% completion of silly sounds Stroop (SSS), and 47% completion of GNG. Because RAs were encouraged to administer tasks in a preset order (SC→WM span→SSS→IS→GNG), individual task completion rates probably reflect a combination of child fatigue and task difficulty. Analyses with IRT (below) provide a formal evaluation of task difficulty.

Dimensionality of Individual EF Tasks

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

SC. A one-factor model fit the 12 items from the SC task poorly, χ²(30, N = 884) = 356.5, p < .0001, CFI = .84, RMSEA = .11. In contrast, a bifactor model, which introduced method factors for car and boat items, along with a general factor,
### Table 1
#### Sample Description at 36-Month Visit

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Total</th>
<th>Completed</th>
<th>No-Opp</th>
<th>Not-Comp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>M (SD)</td>
<td>%</td>
<td>M</td>
</tr>
<tr>
<td>State residence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>40</td>
<td>41</td>
<td>17</td>
<td>51</td>
</tr>
<tr>
<td>NC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC (AA)</td>
<td>41</td>
<td>42</td>
<td>44</td>
<td>33</td>
</tr>
<tr>
<td>TC (AA)</td>
<td>42</td>
<td>43</td>
<td>46</td>
<td>35</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC (female)</td>
<td>99</td>
<td>99</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>PC (male)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>TC (female)</td>
<td>49</td>
<td>53</td>
<td>56</td>
<td>23</td>
</tr>
<tr>
<td>TC (male)</td>
<td>51</td>
<td>47</td>
<td>44</td>
<td>77</td>
</tr>
<tr>
<td>ADHD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk group</td>
<td>17</td>
<td>17</td>
<td>15</td>
<td>22</td>
</tr>
<tr>
<td>PC marital status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>57</td>
<td>58</td>
<td>70</td>
<td>43</td>
</tr>
<tr>
<td>Not married</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC (years)</td>
<td>29.4</td>
<td>29.6</td>
<td>28.9</td>
<td>27.5</td>
</tr>
<tr>
<td>TC (months)</td>
<td>37.1</td>
<td>37.0</td>
<td>38.6</td>
<td>36.5</td>
</tr>
<tr>
<td>PC education (years)</td>
<td>13.0</td>
<td>13.0</td>
<td>13.1</td>
<td>12.4</td>
</tr>
<tr>
<td>Household income and needs</td>
<td>1.9</td>
<td>1.9</td>
<td>2.3</td>
<td>1.3</td>
</tr>
<tr>
<td>TC–ADHD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent report</td>
<td>4.2</td>
<td>4.0</td>
<td>3.8</td>
<td>5.7</td>
</tr>
<tr>
<td>Childcare report</td>
<td>0.8</td>
<td>0.8</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>RA report</td>
<td>0.7</td>
<td>0.6</td>
<td>0.6</td>
<td>1.3</td>
</tr>
<tr>
<td>TC–WPPSI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block Design</td>
<td>90.4</td>
<td>91.6</td>
<td>90.3</td>
<td>76.8</td>
</tr>
<tr>
<td>Receptive Vocabulary</td>
<td>98.9</td>
<td>100.2</td>
<td>97.8</td>
<td>83.4</td>
</tr>
<tr>
<td>FIQ estimate</td>
<td>93.6</td>
<td>95.1</td>
<td>92.9</td>
<td>75.9</td>
</tr>
</tbody>
</table>

**Note.** With the exception of the childcare report of ADHD, for which $N = 455$, total $N = 1,054$–1,123; completed $n = 975$, no-opp $n = 54$, and not-comp $n = 94$. EF = executive function; PA = Pennsylvania; NC = North Carolina; AA = African American; PC = primary caregiver; TC = target child; RA = Research Assistant; ADHD = attention-deficit/hyperactivity disorder; WPPSI = Weschler Preschool and Primary Scales of Intelligence; FIQ = Full scale intelligence; Completed = Children who completed one or more EF tasks; Not-Comp = Children who were unable or unwilling to complete one or more EF tasks; No-Opp = Children who were not given an opportunity to complete.

fit the data well, $\chi^2(36, N = 884) = 65.3, p = .002, \text{CFI} = .99, \text{RMSEA} = .03$.

SSS. A one-factor model fit the 17 items from SSS task poorly, $\chi^2(30, N = 479) = 657.1, p < .0001, \text{CFI} = .71, \text{RMSEA} = .21$. In contrast, a bifactor model, which introduced method factors for dog and cat items along with a general factor, fit the data fairly well, $\chi^2(55, N = 479) = 161.1, p < .001, \text{CFI} = .95, \text{RMSEA} = .06$.

**Animal GNG.** A one-factor model fit the seven no-go items from the GNG task well, $\chi^2(11, N = 444) = 60.0, p < .001, \text{CFI} = .97, \text{TLI} = .97, \text{RMSEA} = .10$. Examination of the results suggested that the set of factor loadings were substantially enough to warrant collapsing of items on each page, resulting in four ordinal score scores (2 two-item and 2 three-item scores). A one-factor model fit the resulting four item WM scale only modestly well, $\chi^2(2, N = 800) = 27.3, p < .001, \text{CFI} = .97, \text{RMSEA} = .12$. Efforts to further improve model fit for this task were unsuccessful.

### IRT: Item Parameter Estimation and Scoring

Unidimensional and bifactor IRT structures were parameterized with either the two-parameter logistic model (2PLM) or Samejima’s (1969) graded response model. The 2PLM was applied to four of the five tasks with dichotomous responses (i.e., SC, SSS, GNG, IS). The graded response model was applied to the WM task because, as described above, the item dependencies between dichotomous responses were collapsed into ordinal scores.
SC. All SC items were at least moderately related to theta (i.e., latent ability underlying performance on SC). Item slopes varied widely (0.74 to 0.79). Consistent with its inclusion as an easy task for young children, all of the difficulty (i.e., $b$) parameters for SC items were less than zero (indicating that they would likely be passed by a child of average ability).

SSS. One item (Number 35) was omitted because it was not significantly related to the theta ($a = .06, SE = .07$). Despite the omission of this item, the bifactor model still fit the data well CFA, $\chi^2(49, N = 479) = 144.5, p < .001$, CFI $=.96$, RMSEA $=.06$. Similar to the SC scale, item slopes varied widely (0.43 to 3.75). Consistent with its inclusion as a moderately difficult task for young children, the majority of difficulty (i.e., $b$) parameters for SSS items were greater than 0 (indicating that they would likely not be passed by a child of average ability).

GNG. Unlike other tasks, the slopes of PG items could be constrained to be equal without degrading model fit (i.e., consistent with a one-parameter logistic model). The item set was strongly related to the construct ($a = 2.19$). Moreover, items became more difficult for children the longer the task continued, irrespective of the number of go trials that preceded no-go trials. Similar to the SC scale, the GNG scale was a relatively easy task for children to complete (item difficulties are primarily less than 0). These results suggest that it was likely child fatigue (the GNG task was the last to be administered in the preset order), not task difficulty, that accounted for the low completion rate for this task.

IS. One item (Number 14) was omitted due to estimation problems (impossibly large standard error for the discrimination parameter). Despite the omission of this item, the bifactor model still fit the data well, $\chi^2(41, N = 846) = 191.8, p < .001$, CFI $=.94$, RMSEA $=.07$. Item slopes varied widely (0.54 to 2.52). Consistent with their inclusion as a moderately difficult task for young children, most of the difficulty (i.e., $b$) parameters for SSS items were between 0 and 1.

WM span. Initial estimation problems (inadmissible standard errors for Item 1) were encountered while estimating WM unidimensional graded response model parameters. To circumvent these problems, the slopes of the first two items were constrained to be equal, and the slopes of the last two items were constrained to be equal. While these constraints did lead to a significant decrement in fit when evaluated with CFA methods, $\chi^2(2) = 23.6, p < .001$, the overall model fit was relatively unchanged, $\chi^2(2, N = 800) = 49.9, p < .001$, CFI $=.95$, RMSEA $=.12$. Unexpectedly, the slope of the first set of two items (i.e., the simpler items) was more discriminating (informative of the underlying construct) than was the slope of the latter set of two items ($b = 3.11$ vs. 1.03). Consistent with its inclusion as a moderately difficult task for young children, all of the difficulty (i.e., $b$) parameters for WM items were between 0 and 3.

Dimensionality of the EF Task Battery

In the process of item parameter estimation described above, IRT-based scores (i.e., EAP estimates of task performance) were computed to reflect each child’s performance on each task. A CFA model with robust full information maximum likelihood was used to test the dimensionality of the task battery. The use of full information maximum likelihood estimation ensured that all 975 children, who completed at least one task, were included in this analysis. A one-factor model fit the task scores well, $\chi^2(5, N = 975) = 3.5, p = .62$, CFI $= 1.0$, RMSEA $= .00$. The latent variance was significant, indicating interindividual differences on task performance. The factor loadings for four of the five tasks were significantly different than 0, with the factor loading of the fifth task, GNG, being marginally significant ($p = .14$). The $R^2$ values for individual task (EAP) scores were highly discrepant, with the single EF factor explaining 1%–48% of the variation in each task (GNG: $R^2 = .01$; SSS: $R^2 = .08$; operation span: $R^2 = .14$; SC: $R^2 = .21$; IS: $R^2 = .48$).

The excellent fit of the one-factor model suggested that testing the fit of two-factor model was likely unnecessary. Nonetheless, given the uncertainty regarding the dimensionality of EF in early childhood, a two-factor model was estimated in which the three tasks that putatively measured inhibitory control (SC, GNG, SSS) loaded separately from a factor defined by the attention shifting and WM tasks. This two-factor model fit the task scores well, $\chi^2(4, N = 975) = 2.4, p = .66$, CFI $= 1.0$, RMSEA $= .00$. Both latent variances were significant, and the factors were positively correlated ($\varphi = .81, p < .001$). The factor loadings for four of the five tasks were significantly different than 0, with the factor loading of the fifth task (GNG) now approaching significance ($p = .07$). The $R^2$ values for most tasks improved very modestly (GNG: $R^2 = .02$; SSS: $R^2 = .09$; operation span: $R^2 = .14$; SC: $R^2 = .29$; IS: $R^2 = .52$). Not surprisingly, the two-factor model did not provide a statistically significant improvement in model fit relative to the one-factor model, $\Delta \chi^2(1) = 1.1, p = .30$, as indicated by a comparison appropriate for model comparisons involving robust maximum likelihood estimation (Satorra & Bentler, 1999).

Test Information and Reliability

A byproduct of IRT estimation is the ability to plot test information curves (TICs), which graphically represent the quality of measurement provided by a task across the range of latent ability that underlies test takers’ performance (theta in IRT parlance). The higher the TIC is at a given level of the construct, the smaller the standard error of measurement. Hence, TICs are directly related to the reliability of each task. Given that all five EF tasks conform to a single dimension, TICs from all five tasks were plotted along a single dimension (see Figure 1). Although the relative height of TICs across tasks are not comparable (due to the fact that tasks have different numbers of items contributing to scores), the spread of TICs is interpretable. Specifically, the battery appears to successfully measure a relatively wide range of EF ability level, with some tasks (i.e., SC, GNG, and IS) being most informative of ability among low performing children and other tasks (i.e., SSS, WM) being most informative of ability among higher performing children. An alternative way to represent the reliability of each task as a function of child ability level is to compute the standard average reliability of each EF task as a function of child ability level. Table 2 summarizes this information, given the reliability of each IRT-based task score across a wide range of ability level.

---

1 The unexpected change in degrees of freedom between this model and that previously reported is due to the use of the WLSMV estimator.

2 Although large slope estimates (e.g., 3.75) may be indicative of local dependence, subsequent analyses ruled out this explanation.
(standard normal scores ranging from −3 to 3) in increments of 1 standard deviation. With the exception of the WM task, the reliability of each task is the highest (and hence the standard error of measurement is the lowest) for children whose ability is between −1 and 1 standard deviations from the sample mean. Due to the high difficulty of the WM task, scores for that task are most informative for children whose ability is between 0 and 2 standard deviations above the sample mean. Consistent with Figure 1, the tasks do a progressively poorer job (reliability decreases; the standard error of measurement increases) of measuring EF for children at extremely low or high ability levels.

**Criterion Validity**

A final CFA model was estimated to establish the criterion validity of the EF task battery. One latent variable represented children’s performance on the EF task battery. The remaining two latent variables represented criterion measures: multiple informant (parent, day care provider, RA) ratings of ADHD behaviors and estimated intellectual functioning (based on performance on Block Design and Receptive Vocabulary subtests of the WPPSI–III). Bivariate correlations among the indicators of EF, ADHD, and intelligence latent factors are summarized in Table 3. Individual EF tasks were moderately, positively correlated with scores from the WPPSI–III subtests (rs = .10 to .46). Individual EF tasks were modestly, negatively correlated with parent and RA ratings (rs = −.05 to −.37) but closer to 0 for childcare provider ratings of hyperactivity (rs = −.14 to .06).

This criterion validity CFA model used data from the 1,122 children who participated in the 36 month assessment and who had data on at least one indicator of any of the three latent variables. Although some of the cases that contributed data to this model did not complete any EF tasks, they were included in order to minimize the potential for selection bias. This three-factor model fit the observed data well, $\chi^2(32) = 36.0, p = .29$, CFI = 1.0, RMSEA = .01. The latent variances for all three factors were significant ($ps < .001$), and the factor loadings for all indicators (including the loading of GNG) were statistically different from 0 ($ps < .001$, except the GNG task in which $p = .03$). In support of hypotheses, the EF latent variable was strongly, negatively correlated with ADHD ($\varphi = -.71, p < .001$) and strongly positively correlated with IQ ($\varphi = .94, p < .001$). The ADHD and IQ latent variables were also strongly, negatively correlated with each other ($\varphi = -.72, p < .001$). These latent correlations are appreciably larger in magnitude (lsll of .72 and .94) than are the bivariate correlations reported among indicators of EF, ADHD, and IQ (lsrl from .10 to .40). This discrepancy is attributable to the fact that latent correlations are based solely on true score variation that is shared across indicators of each latent construct, whereas bivariate correlations include a combination of true score variation, unique (task specific systematic) variation, and measurement error.

As a final check of criterion validity, children were classified into low and high risk ADHD groups, in which risk was defined as parent reports of six or more inattentive and/or hyperactive–impulsive symptoms (loosely approximating Diagnostic and Statistical Manual of Mental Disorders (4th edition)/criteria for ADHD). Of the 1,096 participants who had complete data on the ADHD rating scale, 17% met this criterion for risk. ADHD risk status did not significantly differ as a function of the availability of EF data, $\chi^2(2, N = 1096) = 1.8, p = .42$. (ADHD risk: 17% among those who completed at least one EF task, 15% among those who did not have an opportunity to complete EF tasks, 22% among those who were unable to complete any EF tasks). Children in the high ADHD risk group exhibited an average of 12.5 symptoms ($M = 5.6$ inattentive; $M = 6.9$ hyperactive–impulsive), whereas children in the low ADHD risk group exhibited an average of 2.4 symptoms ($M = 0.8$ inattentive; $M = 1.6$ hyperactive–impulsive).

A logistic regression model was estimated in which ADHD risk status was regressed on each child’s estimated EF ability (i.e., factor score estimate of EF ability from the one-factor CFA model of EF tasks). Factor scores were standardized to have $M = 0$ and $SD = 1$ (high risk group $M = −0.31$ versus low risk group $M = 0.07$; Cohen’s $d = .38$). Performance on EF tasks predicted ADHD risk status, likelihood ratio: $\chi^2(1, N = 964) = 27.7, p < .0001$. Higher scores on the EF battery were associated with reduced likelihood of membership in the ADHD risk group (odds ratio = .62, 95% CI [.50–.76]). Nonetheless, the strength of the association was modest. The model (maximum rescaled) $R^2 = .05$ and the

---

**Table 2**

| Reliability Estimates in 1 Standard Deviation Unit Increments From −3 to 3 Standard Deviations Around the Mean Level of EF for the SC, SSS, GNG, IS, and WM Tasks |
|---|---|---|---|---|---|
| -- | SC | SSS | GNG | IS | WM span |
| −3 to −2 | .51 | .28 | .23 | .41 | .04 |
| −2 to −1 | .77 | .53 | .66 | .69 | .11 |
| −1 to 0 | .88 | .81 | .87 | .85 | .37 |
| 0 to 1 | .78 | .90 | .82 | .84 | .78 |
| 1 to 2 | .47 | .84 | .46 | .68 | .80 |
| 2 to 3 | .17 | .63 | .09 | .43 | .50 |

*Note.* EF = executive function; SC = spatial conflict; SSS = silly sound Stroop; GNG = go/no-go; IS = item selection; and WM = working memory.
concordance index (c) = .60 (where values range from 0.50–1.0, reflecting completely chance to perfect association). The mean predicted probability in the total sample was .17 (range .07–.33).3 Classifying cases based on predicted probabilities of .24 (which corresponded to the 90th percentile) resulted in an overall correct classification rate of 78% (752/964), with good specificity (.84 (725/860) but poor sensitivity = .17 (27/162) and positive predictive power = .26 (27/104). Hence, whereas strong performance on the EF battery was informative of low risk for ADHD, poor performance on the EF battery was less informative of high risk for ADHD.

**Discussion**

Given the relation of EFs to a number of aspects of child development—including self-regulation, mental development, 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 development of EF in early childhood, including the identification of naturally occurring experiences, as well as experimental interventions, which promote competence and resilience in children at risk for school failure and early developing psychopathology (Blair, Zelazo, & Greenberg, 2005). With these goals in mind, this study reported the psychometric properties, the factor structure, and the criterion validity of a newly developed battery of tasks designed to measure EF at age 3 years with a population-based sample of children residing in low-income and nonurban communities.

Of the 1,069 children who were given an opportunity to complete the EF battery (i.e., those for whom an in-home visit at 36 months was completed), 91% (N = 975) of children were able to complete at least one task. To the best of our knowledge, this is the first study that has presented EF tasks to children in a representative sample selected at birth with no exclusions and thus included children with significant developmental disabilities. The fact that 91% of children could complete at least one task demonstrates that direct child assessments of EF in the context of large-scale studies are feasible. Extensive pilot testing with low-income families underscored our impression that tasks should be presented in a structured manner, with as minimal language demands as possible. Despite these best efforts, there continued to be 9% of the children who were unable to complete any of our tasks. Children unable to complete any tasks were markedly different from the larger sample with respect to their cognitive functioning, as well as with respect to their behavioral functioning and home environments. Dramatically simplified tasks may be necessary to accurately assess the EF abilities for these riskiest of children.

Most individual EF tasks were better represented with bifactor, not unidimensional, models. Bifactor models represent a strategy for taking into account residual correlations between items that remain even after modeling item interpolations due to a shared common factor. For example, children found it more difficult to “meow” when seeing pictures of dogs than they did to “bark” when seeing pictures of cats. This was not anticipated. The necessity of introducing method factors for most of our tasks underscored the idiosyncratic ways in which 3-year-old children responded to some tasks. Our modeling approach purged this idiosyncratic (nuisance) variation from the computation of EF task scores.

When scores for the entire battery were considered together, a unidimensional model was found to fit the data extremely well. This result is consistent with that of three previous studies in which the dimensionality of EF in young children was examined (Carlson, Mandell, & Williams, 2004; Hughes & Ensor, 2007; Wiebe, Espy, & Charak, 2008). This study adds to a growing body of work that has demonstrated that although EF abilities may be best conceptualized as being multidimensional in older children and adults, they are better conceptualized as being unidimensional (undifferentiated) in early childhood. One caveat is that in studies of older children and adults a large number of tasks were administered, which provides a stronger test of dimensionality. This strategy presents unique challenges for young children (test burden) that could not be easily addressed within the context of the current study.

Although numerous tasks that were purported to measure EF in early childhood have been developed, most widely used tasks have not undergone formal psychometric evaluations and provide no

### Table 3

**Pairwise Bivariate Correlations Between IRT Scores for Each Executive Function Task and Indicators of ADHD and Intelligence**

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Spatial conflict</td>
<td>—</td>
<td>.17</td>
<td>.12</td>
<td>.35</td>
<td>.19</td>
<td>.27</td>
<td>.30</td>
<td>.17</td>
<td>.14</td>
<td>.30</td>
</tr>
<tr>
<td>2. Silly sound Stroop</td>
<td>.13</td>
<td>—</td>
<td>.11</td>
<td>.22</td>
<td>.11</td>
<td>.13</td>
<td>.14</td>
<td>.11</td>
<td>.01</td>
<td>.19</td>
</tr>
<tr>
<td>3. Go/no-go</td>
<td>.04</td>
<td>.04</td>
<td>—</td>
<td>.08</td>
<td>.07</td>
<td>.09</td>
<td>.16</td>
<td>.12</td>
<td>.02</td>
<td>.17</td>
</tr>
<tr>
<td>4. Item selection</td>
<td>.31</td>
<td>.19</td>
<td>.02</td>
<td>—</td>
<td>.30</td>
<td>.45</td>
<td>.49</td>
<td>.20</td>
<td>.19</td>
<td>.40</td>
</tr>
<tr>
<td>5. Working memory span</td>
<td>.18</td>
<td>.05</td>
<td>.02</td>
<td>.29</td>
<td>—</td>
<td>.29</td>
<td>.35</td>
<td>.10</td>
<td>.01</td>
<td>.25</td>
</tr>
<tr>
<td>6. Block Design</td>
<td>.29</td>
<td>.10</td>
<td>.05</td>
<td>.38</td>
<td>.28</td>
<td>—</td>
<td>.48</td>
<td>.17</td>
<td>.16</td>
<td>.42</td>
</tr>
<tr>
<td>7. Receptive Vocabulary</td>
<td>.29</td>
<td>.10</td>
<td>.10</td>
<td>.46</td>
<td>.34</td>
<td>.46</td>
<td>—</td>
<td>.21</td>
<td>.19</td>
<td>.40</td>
</tr>
<tr>
<td>8. ADHD–PC</td>
<td>.12</td>
<td>.06</td>
<td>.05</td>
<td>.12</td>
<td>.13</td>
<td>.17</td>
<td>.20</td>
<td>—</td>
<td>.26</td>
<td>.33</td>
</tr>
<tr>
<td>9. ADHD–CC</td>
<td>.11</td>
<td>.02</td>
<td>.06</td>
<td>.14</td>
<td>.04</td>
<td>.20</td>
<td>.18</td>
<td>.22</td>
<td>—</td>
<td>.30</td>
</tr>
<tr>
<td>10. ADHD–RA</td>
<td>.27</td>
<td>.11</td>
<td>.09</td>
<td>.31</td>
<td>.18</td>
<td>.37</td>
<td>.36</td>
<td>.31</td>
<td>.28</td>
<td>—</td>
</tr>
</tbody>
</table>

**Note.** N = 201–1,121. Values above and below the diagonal are weighted and unweighted correlations, respectively. IRT = item response theory; ADHD = attention-deficit/hyperactivity disorder; PC = primary caregiver rating; CC = Childcare provider rating; RA = Research assistant rating.

---

3 Because predicted probabilities are not available with PROC SURVEYLOGISTIC in SAS (version 9.1), all predicted probabilities and classification information was based on results from the model reestimated with PROC LOGISTIC. Although the standard errors from this model are inappropriate, the predicted probabilities and associated inferences about classification are unaffected.
reliability data (the NEPSY neuropsychological battery—which includes some assessment of EF in addition to assessments of language, visual-spatial processing, sensorimotor function, and memory and learning—and the aforementioned shape school task are two notable exceptions; Espy, Bull, Martin, & Stroup, 2006; Korkman, Kirk, & Kemp, 1998). Our IRT-based approach permits a consideration of how reliability differs as a function of ability level, which is not possible from the perspective of classic test theory. Specifically, TICs revealed that the current battery provides more reliable measurement of EF ability for 3-year-old children within 1 standard deviation of the sample mean than it does for children who are extremely high or low functioning in EF. The routine presentation of TICs for measures of EF in early childhood would facilitate the selection of tasks to closely match the characteristics of children under study. For example, if we were designing a new study that included young children or children who were otherwise expected to exhibit relatively immature EF abilities (e.g., children living in poverty; children at risk for autism), we would include our SC task. In contrast, if we were designing a new study that focused on older preschoolers or children who were otherwise expected to exhibit relatively advanced EF abilities (e.g., children residing in resource rich homes: cognitively precocious children), we would include our WM span task. The matching of specific tasks to specific ability levels is probably more realistic than identifying tasks that work extremely well for children of all ability levels—though that is the standard to which task developers should continue to strive.

Criterion validity analyses demonstrated that children who performed best on the EF battery were rated by multiple adult informants (parents, childcare providers, and RAs who observed children during home visits) as exhibiting the lowest levels of ADHD behaviors. When risk for ADHD was considered dichotomously, strong performance on the EF battery was associated with low risk for ADHD; however, the converse was not true—poor performance on the EF battery was not associated with high risk of ADHD. These results are consistent with the notions that although a subset of ADHD youths appears to exhibit impairments in EF, poor performance on EF tasks is neither necessary nor sufficient for diagnosis (Grodzinsky & Barkley, 1997; Nigg et al., 2005).

Also consistent with our hypotheses, children who performed best on EF tasks also performed best on two subtests of the WPPSI–III, which at age 3 years can be use to approximate full scale IQ. The exceptionally strong latent correlation between task performance on our EF battery and select WPPSI–III tasks is consistent with a well-established literature that although EF and intelligence, particularly crystallized intelligence, are clearly differentiated in a number of literatures (Blair, 2006), the WM aspect of EF is highly related to measures of intelligence in adults (Friedman et al., 2006; Gustafsson, 1984; Kane et al., 2005). Our results are the first to demonstrate that a similarly strong association exists in early childhood.

This study was characterized by at least three strengths. First, the battery was tested with a representative sample of 3-year-old children, with oversampling for low-income and (in North Carolina) African American families. Working with a large number of children who resided in low-income families increased our confidence that the tasks would work (i.e., were comprehended and tolerated) with children who have been underrepresented in research on EF in early childhood. Moreover, the combined use of an explicit sampling design and an analysis plan that acknowledged the stratified sampling plan with oversampling permitted us to generalize our results to a broader population of children than had previous studies that relied on convenience samples who are motivated to participate in university research. Second, the tasks were developed to facilitate standard administration by lay interviewers who did not have expertise in EF. Relatively high rates of completion demonstrate the feasibility of direct child assessments in the context of large-scale studies. Third, the novel application of IRT methods demonstrated how the reliability of EF tasks varies as a function of child ability level. In the future, it may be beneficial to explicitly develop EF tasks that are optimized to the assessment of children with specific ability profiles.

This study is also characterized by at least four weaknesses. First, 9% of children were unable to complete any EF tasks. Moreover, even among those who could complete tasks, many could not complete the entire battery. Given completion of the battery was embedded within a larger visit that made other demands of children and that necessitated the administration of tasks in a fixed order, this study is not well positioned to determine the relative contributions of overall visit fatigue versus task difficulty in explaining low rates of completion for some tasks. Second, despite the ecological validity of administering EF tasks in children’s homes, many households were characterized by nonoptimal test conditions (e.g., frequent interruptions from others in the household, poor lighting, high levels of ambient sound), which likely differentially under-mined children’s task performance. Third, the FLP sample is representative of 2 three-county areas that are nonmetropolitan and that are characterized by moderately high poverty. The results of this study are in no way nationally representative of all 3-year-old children. Fourth, it will be important to consider a broader set of criterion (both behavioral and performance-based) measures in order to continue to validate this new battery.

Despite an explosion of research on children’s self-regulation in early childhood, the field continues to be dependent on tasks that have not been subjected to rigorous psychometric evaluation. Moreover, given a central assumption that early childhood is characterized by rapid developmental onset of EF abilities, it will be imperative to develop scalable instruments that facilitate inferences about interindividual differences in intra-individual change in EF across ages 3–5 years. This study represents our initial step in this work. The development of psychometrically sound, scalable measurement tools, which facilitate the study of interindividual differences in EF ability and which can be used in the context of large-scale studies, will dramatically improve the scientific study of EF in early childhood.

References


function, and false-belief understanding to emerging math and literacy ability in kindergarten. Child Development, 78, 647–663.


Received August 26, 2009
Revision received November 26, 2009
Accepted November 29, 2009

317MEASUREMENT OF EXECUTIVE FUNCTION IN 3-YEAR-OLDS