Published in final edited form as:

Psychol Assess. 2013 June ; 25(2): 664–670. doi:10.1037/a0031747.

Measuring Executive Function in Early Childhood: A Focus on Maximal Reliability and the Derivation of Short Forms

Michael T. Willoughby,

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

Jolynn Pek, and

Department of Psychology, York University

Clancy B. Blair

Department of Applied Psychology, New York University.

Family Life Project Investigators

Abstract

This study provided estimates of the maximal reliability of a newly developed battery of executive function (EF) tasks for use in early childhood. In addition, it demonstrated how changes in maximal reliability can inform the selection of different “short forms” of the battery, depending on child age. Participants include children from the Family Life Project—a prospective longitudinal study ($N = 1292$) of families who were over-sampled from low income and African American families at the birth of a new child—at age 3, 4, and 5 year assessments. Results indicated that the EF battery had reasonably good maximal reliability ($\rho = .73$, 95% CI = .69 - .76) in a mixed-age sample that included children who were randomly selected from the age 3, 4, and 5-year assessments. In contrast, the maximal reliability of the battery ranged from poor to modest for within-age samples (ρ s = .47 [95% CI = .37 - .52], .62 [95% CI = .57 - .66], and .61 [95% CI = .55 - .66] at ages 3, 4, and 5 years, respectively). Although the derivation of a three-task “short form” of the battery always resulted in statistically significant decrements in maximal reliability, in some cases, the relative decrement in maximal reliability was quite modest and may be tolerable given the time savings and potential reduction in participant burden. Results are discussed with respect to the benefits of using maximal reliability to both evaluate task batteries and derive short forms, as well as how a focus on maximal reliability informs ongoing questions about the measurement and conceptualization of EF in early childhood.

Keywords

Executive Function; Early Childhood; Maximum Reliability

Correspondence should be sent to Michael Willoughby, FPG Child Development Institute, UNC-CH, Campus Box 8185, 521 South Greensboro Street, Carrboro, NC 27510. willoughby@unc.edu..

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/pas

EF is an “umbrella” term and serves as a general referent to a set of cognitive abilities that are drawn upon when encountering novel problems or goal oriented activities (see Anderson, 2008). Recent research has focused on the developmental course, dimensionality, and malleability of EF. Findings have indicated that EFs follow a protracted developmental course (e.g., Romine & Reynolds, 2005), show evidence of both unity and diversity (e.g., Garon, Bryson, & Smith, 2008; Miyake & Friedman, 2012), and are malleable, showing deficits in response to stress (e.g., Arnsten, 1998, 2000) and enhancements in response to structured activities (e.g., Bryck & Fisher, 2012).

Confirmatory factor analytic (CFA) models have been advocated as a method of choice for empirically defining the construct of EF, as well as for testing questions about its dimensionality (Miyake et al., 2000; Wiebe, Espy, & Charak, 2008). An emerging consensus has indicated that whereas EF is an unidimensional construct in early childhood, it is multidimensional in late middle childhood and beyond (see e.g., Shing, Lindenberger, Diamond, Li, & Davidson, 2010). Hence, although there is no strong theory about the dimensionality of EF, the empirical literature provides support for the notion of unity *and* diversity in EF (Garon et al., 2008; Miyake & Friedman, 2012).

In the case of early childhood, not only have single factor models of EF tended to provide the best *relative* fit to the data (but see Schoemaker et al, 2012 for a recent exception), they have also provided outstanding *global* model fit. Many researchers may have erroneously assumed that excellent global model fit was synonymous with excellent measurement. Paradoxically, CFA models with *poorer* quality measurement often have *better* global model fit than do comparable models that are characterized by better quality measurement (Hancock & Mueller, 2011). As noted in this *Journal* and elsewhere, whereas single factor models provided excellent global model fit to our newly developed battery of EF tasks, the latent variable only explained 25-50% of the observed variation in any given task (Willoughby & Blair, 2011; Willoughby, Blair, Wirth, Greenberg, & Investigators, 2010, 2012). Similar results were evident in other studies involving young children (Hughes, Ensor, Wilson, & Graham, 2010; Wiebe et al., 2008; Wiebe et al., 2011) and adults (Ettenhofer, Hambrick, & Abeles, 2006; Miyake et al., 2000). Hence, this phenomenon may not be specific to our battery or to a given age period.

Maximal reliability provides a quantitative index of the quality of measurement of a latent variable from a given set of indicators (Hancock & Mueller, 2001; Raykov, 2004; Raykov & Hancock, 2005). The first objective of this study was to provide estimates of the maximal reliability (ρ) of a recently developed battery of EF tasks for use in early childhood. Given that most studies of EF in early childhood include children who span from 3-5 years, we initially provide the maximal reliability of latent EF ability as measured by the battery for a sample of children who were sampled across the age 3, 4, and 5-year assessments. However, because children in this study completed the battery at multiple ages, we also provide separate estimates of the maximal reliability of the battery at each assessment. Given appreciable mean level changes in EF performance on this battery between age 3-5 years (Willoughby, Wirth, Blair, & Investigators, 2012), we expected that maximal reliability would be larger in the mixed-age sample than the within-age sample. That is, given age-

based variation in EF tasks, we expected stronger correlations between EF tasks in the mixed versus within-age sample.

Since developing our EF task battery, a ubiquitous question is whether the battery can be reduced in length and, if so, which subset of tasks is recommended as an optimal “short form” measure of EF. Following Raykov and Hancock (2005), we considered changes in the maximal reliability as a principled approach for addressing this question. Because the maximal reliability of a latent construct, as defined here, cannot improve with the removal of indicators, a comparison of the difference in maximal reliability between the full and partial subset of tasks provides an empirical estimate of the “cost” that results from using a subset of tasks from the total battery. Hence, the secondary objective of this study was to determine the maximal reliability of all possible 3-task combinations of the battery. Whereas five tasks were administered at the age 3 year assessment, six tasks were administered at the age 4 and 5 year assessments. Given that each task takes approximately five minutes to complete, reducing the battery to three tasks would result in a 50-60% reduction in direct child assessment time. In many settings, especially large scale studies, this may dictate whether EF is assessed at all.

Methods

Participants

The Family Life Project (FLP) is a prospective longitudinal study of a representative sample of $N=1292$ children who reside in two of the four major geographical areas of the United States with high poverty rates. Full details regarding the sampling and study design, as well as a description of families and communities is provided elsewhere (Vernon-Feagans, Cox, & Investigators, in press). This study utilized data from the age 3, 4, and 5 year assessments.

Of the $N=1123$ (87% of the total sample) families who participated in the age 3 assessment, $N=973$ children completed at least one EF task ($M = 3.5$, $SD = 1.3$ of five possible tasks). Of the $N=1066$ (83% of the total sample) families who participated in the age 4 assessment, $N=1009$ children completed at least one EF task ($M = 5.5$, $SD = 1.0$ of six possible tasks). Of the $N=1099$ (85% of the total sample) families who participated in the age 5 assessment, $N=1036$ children completed at least one EF task ($M = 5.8$, $SD = 0.7$ of six possible tasks). The discrepancies between the number of families who participated in a given visit and the number of children who completed EF tasks was primarily related to some visits being completed by phone (i.e., for families who had moved > 200 miles from the study area), in which case direct assessments with children were not possible, as well as some children being unable to complete any EF tasks (i.e., 9% and 1% of children who were given an opportunity to complete EF tasks were unable to do so at ages 3 and 5 years, respectively). A total of $N=865$ children (67% of the total sample) participated in EF testing at all three assessments, $N = 171$ children (13%) at two of three assessments, $N = 87$ children (7%) at one of three assessments, and $N = 171$ (13%) at none of the assessments. Children who did not complete any EF assessments ($N = 171$) did not differ from children who completed one or more assessments ($N=1121$) with respect to state of residence (36% vs. 41% residing in PA, respectively, $p = .26$), being recruited in the low income stratum (77% vs. 78% poor, p

= .75), race (37% vs. 43% African American, $p = .15$), or gender (56% vs. 50% male, $p = .19$).

We also created cross-sectional sample of mixed-age children who completed at least one EF assessment ($N = 1121$). This sample was created by sampling observations from across assessments. Specifically, for children who only participated in a single EF visit, we sampled their available data from that assessment. For children who participated in two or three assessments, we randomly sampled from which assessment their data was drawn. The sampling scheme ensured equal representation of child observations across the 3-5 year assessments; $N=373$ (33%), $N=374$ (33%) and $N=374$ (33%) child observations were drawn from the age 3, 4 and 5 year assessments, respectively. This mixed-age sample more closely represented extant studies of EF in early childhood, which typically involve participants of varying ages.

Procedures

Families participated in home visits when study children were 3 (2 visits), 4 (1 visit) and 5 (1 visit) years old. On average, each home visit took approximately two hours to complete, of which children spent approximately 30-45 minutes (including practice, administration, and breaks between tasks) completing EF tasks. Full details on the testing procedure were described in previous manuscripts in this *Journal* (Willoughby et al., 2010; Willoughby, Blair, et al., 2012).

Measures

The EF battery consisted of six tasks. Given space constraints, we provide an abbreviated description of tasks below.

Working Memory Span (WMS)—This span-like task is based upon principles described by Engle, Kane and collaborators (e.g., Kane & Engle, 2003). The task requires children to perform the operation of naming and holding in mind two pieces of information simultaneously and to activate one of them (i.e., animal name) while overcoming interference occurring from the other (i.e., color name). Based on the results of pilot testing (focused on test burden), only the first 11 items were administered at the 3-year assessment, whereas all 19 items were administered at the 4 and 5-year assessments.

Pick the Picture Game (PTP)—This is a Self-Ordered Pointing task (Cragg & Nation, 2007; Petrides & Milner, 1982). This task requires working memory because children have to remember which pictures in each item set they have already touched. The PTP was too difficult for many 3 year olds and only administered at the 4- and 5-year assessments.

Silly Sounds Stroop (SSS)—This task was derived from the Day-Night task developed by Gerstadt, Hong, and Diamond (1994). This task requires inhibitory control as children have to inhibit the tendency to associate bark and meow sounds with dogs and cats, respectively.

Spatial Conflict (SC)—This is a Simon task similar to that used by Gerardi-Caulton (2000). Whereas children respond to the initial set of items by touching a response card in the same position as the stimuli (e.g., the stimuli is presented on the left side of the test booklet and the correct response requires that the child touch the left side of his/her response card), the test items require a contra-lateral response (e.g., the stimuli is presented on the left side of the test booklet and the correct response requires that the child touch the right side of his/her response card; spatial location is no longer informative). The SC was administered at the 3-year assessment.

Spatial Conflict Arrows (SCA)—This task was identical in format to the SC task (above) with the exception that the test stimuli and response cards were altered to make the task slightly more difficult. The SCA was administered at the 4 and 5-year assessments.

Animal Go No-Go (GNG)—This is a standard go no-go task (e.g., Durston et al., 2002). The task includes varying numbers of go trials prior to each no-go trial, including, in standard order, 1-go, 3-go, 3-go, 5-go, 1-go, 1-go, and 3-go trials. No-go trials require inhibitory control.

Something's the Same Game (STS)—This task was derived from Jacques and Zelazo's (2001) flexible item selection task. This task requires the child to shift his/her attention from the initial dimension of similarity to a new dimension of similarity.

Scoring EF Tasks

As previously discussed (Willoughby, Wirth, et al., 2012), task scoring was facilitated by drawing a calibration sample of children—all of who were deemed to have high quality data (e.g., data collectors did not report interruptions, children completed multiple tasks)—from across the 3, 4, and 5-year assessments (no child contributed data from more than one assessment). Graded response models were used to score the two tasks with polytomous item response formats (i.e., PTP, WMS), while two-parameter logistic models were used to score the remaining tasks (all of which involved dichotomous items response formats) in the calibration sample. The set of item parameters that was obtained from calibration sample was applied to *all* children's EF data across *all* assessments resulting in a set of item response theory based (i.e., expected a-posteriori [EAP]) scores for each task that were on a common developmental scale. Those scores—which were all normally distributed (i.e., all skew and kurtosis < |1.2|)—served as the indicators for the confirmatory factor models that were estimated here.

Analytic Strategy

Separate sets of analyses were used with the age 3, 4, and 5 year data, as well as the mixed-age sample (i.e., identical analyses were repeated four times). Each analysis involved estimating a single-factor confirmatory factor analytic (CFA) model using all of the individual EF tasks that were available at a given assessment. This included five tasks at the age 3 year assessment (i.e., SC, SSS, GNG, WMS, STS), six tasks at the age 4 and 5 year assessments (i.e., SCA, SSS, GNG, WMS, STS, PTP), and seven tasks for the mixed age sample (i.e., SC, SCA, SSS*, GNG*, WMS*, STS*, PTP; where tasks marked with ‘*’

were common across all ages/samples). Support for a single factor model for these tasks was previously reported (Willoughby et al., 2010; Willoughby, Blair, et al., 2012; Willoughby, Wirth, et al., 2012).

We computed rho (ρ), an estimate of maximal reliability, for the full set of available tasks at a given age, as well as for up to 20, 3-task combinations (the number of 3-task combinations was smaller for the age 3 data than the age 4, age 5, and mixed age data because one fewer task was administered). Rho was computed following equation 5 of Raykov and Hancock (2005). This included using linear constraints of factor loading and item intercepts to obtain estimates of rho and obtaining bias-corrected bootstrapped standard errors to obtain 95% CIs for rho(s). CFA models used robust maximal likelihood (MLR) estimation as implemented in version 6.1 of *Mplus* software. All CFA models took into account individual probability weights but not stratification variables (used for participant recruitment), as the latter cannot be used in conjunction with the *Mplus* bootstrapping procedure that was essential for the computation of confidence intervals.

Results

Maximal Reliability

Mixed Age Sample—Although we have previously established that a single factor model provided a good fit to the covariance between tasks scores separately at each assessment (ages 3, 4, 5 years), the fit of the one-factor model in the mixed-age sample has not been previously reported. The one-factor CFA model fit the data well in this mixed age sample, $\chi^2(12) = 24.2$, $p = 0.02$, CFI = 0.97, RMSEA = 0.03 (90% CI: .01 - .05). Maximal reliability was computed as $\rho = .73$ (95% CI: .69 - .76) for the full set of seven tasks, as $\rho = .72$ (95% CI: .68 - .75) for the set of six tasks that were administered at the age 4 and 5 year assessments, and $\rho = .62$ (95% CI: .57 - .67) for the set of five tasks that were administered at the age 3 year assessment.

Within Age Samples—Maximal reliability was also calculated at the age 3, 4, and 5 year assessments. Maximal reliability was computed as $\rho = .47$ (95% CI: .37 - .52) for the set of five tasks that were administered at the age 3 year assessment, as $\rho = .62$ (95% CI: .57 - .66) for the set of six tasks that were administered at the age 4 assessment, and as $\rho = .61$ (95% CI: .54 - .66) for the same set of six tasks that were administered at the age 5 year assessment. As expected, the within age estimates of maximal reliability were consistently smaller than the comparable estimates that were obtained in the mixed age sample (e.g., compare $\rho = .47$ vs. .62 for the set of five tasks that was administered at the age 3 year assessment for within and mixed age samples, respectively).

Derivation of Short Forms

Mixed Age Sample—The maximal reliability and corresponding 95% CIs of 20 possible, 3-task combinations was computed (see Figure 1). The dashed lines in Figure 1 represents the maximal reliability of EF as defined by the full set of six tasks ($\rho = .72$). Notice that none of the three-task combinations have maximal reliability values that span this threshold. This indicated that there is a “cost” (i.e., decrement in maximal reliability of the latent

variable of EF) when only a subset of tasks was used. Nonetheless, some 3-task combinations were better than others.

The information depicted in Figure 1 was intended to be descriptive. We computed formal tests of difference in maximal reliability for each three task combination relative to that from the full set of six tasks. These estimates of changes in maximal reliability, which we refer to as delta rho ($\rho = \rho_{\text{full set of tasks}} - \rho_{\text{subset of tasks}}$), are summarized in an online supplementary table. Consistent with Figure 1, in no case did the 95% CI for delta rho include 0. This provided more formal support for the idea that all 3-task short forms of the battery resulted in lower maximal reliability of the construct of EF than was the case for the full subset of tasks.

Within Age Samples—Figure 2 depicts the maximal reliability and corresponding 95% CIs of each 3-task combination at ages 3, 4 and 5 years. Because the PTP task was not administered at the age 3 year assessment, there are no age 3 estimates provided for three-task combinations that involve this task. Moreover, the three-task combinations that include “SCA” should be understood to involve the SC task when they are provided for three year old estimates (i.e., the SC task that was administered at the age 3 year assessment was replaced by the SCA task at the age 4 and 5 year assessments); this convention facilitates a more economical depiction of results. With three exceptions at age 3, none of the three-task combinations have maximal reliability values that span that which is achieved when a full set of tasks is used. Consistent with the mixed age sample results, this indicated that the decision to use fewer tasks results in significant decrements in maximal reliability relative to the use of the full battery. Formal tests of reductions in maximal reliability that result from using short forms of the battery are summarized in an online supplementary table. In no instance did the 95% CI for delta rho include 0.

Discussion

Numerous researchers have emphasized the value of using CFA models to define the construct of EF as that variation that is shared across a set of tasks that require cognitive processes necessary for problem solving and goal-directed efforts. However, the good (often excellent) fit of CFA models to observed EF data should not be confused as evidence that selected tasks necessarily measure the latent variable of EF well. Maximal reliability provides a formal estimate of the quality of measurement of the latent variable of EF. Our results indicated that whereas our battery did a respectable job of measuring the latent construct of EF in a mixed age sample ($\rho \approx .70$), it did only a modest job of measuring the latent construct of EF in samples consisting of primarily 4 or 5 year olds ($\rho \approx .60$) and a relatively poor job in a sample consisting of primarily 3 year olds ($\rho \approx .50$).

The fact that maximal reliability of the mixed age sample exceeded that reported at any single age underscores a subtle point; namely, the measurement properties of EF tasks are related to the between subject variability in task performance. To the extent that EF is measured in a mixed age sample, where mean level differences in performance in age are evident, that battery of tasks will appear to do a better job of measuring the construct of EF than will a comparable set of tasks that are administered to a sample of participants who are

highly similar in age. The use of the mixed age sample was intended to facilitate direct comparisons of the maximal reliability of this battery in this sample with comparable other studies which nearly always involve children of mixed ages. Finally, although the estimates of maximal reliability that have been reported here do not invalidate the previously established criterion validity of this battery (Willoughby et al., 2010; Willoughby, Blair, et al., 2012), they indicate that the magnitude of reported criterion validity could have been improved had the latent variable of EF been measured with greater precision, that is larger maximal reliability (see Penev & Raykov, 2006).

Although maximal reliability has not been routinely reported in previous studies that used CFA methods to define EF, it is easy to compute provided that standardized factor loadings are available (see supplementary appendix). Although the focus of a separate manuscript, we have documented low levels of maximal reliability for latent variable representations of EF in many other studies that used a variety of different tasks for participants ranging from early childhood through old age (Willoughby, Holochwest, Blanton, & Blair, in preparation). To the extent that low maximal reliability is a characteristic of the *construct* of EF and not simply a limitation of specific EF tasks or their use during a specific developmental period, this may indicate the need to re-examine long standing concerns related to the measurement of EF (Rabbitt, 1997).

Given poor to moderate levels of maximal reliability of latent variables of EF, future studies may benefit from administering a larger number of tasks in order to achieve high levels of maximal reliability. This is consistent with conventional wisdom regarding measurement—increasing the number of item on a scale (or here the number of tasks in a battery) is a well-established method for improving reliability (Nunnally & Bernstein, 1994). In the context of research studies, the use of planned missing designs, in which a large battery of tasks is administered but each child only completes a subset of all tasks (Graham, Taylor, Olchowski, & Cumsille, 2006), may serve the dual function of improving the quality of measurement of the latent variable of EF without imposing undue test burden on participants, particularly of young children. Obviously, this is not a solution for improving the measurement of EF in clinical assessment contexts.

A secondary focus of this study was the use of maximal reliability as a method for informing the development of “short forms” of task batteries. Although a focus on task reduction is somewhat at odds with the foregoing discussion indicating the likely need for more not fewer tasks to optimize the reliability of the construct of EF, practically speaking, there are many situations in which researchers have only a short amount of direct child assessment time. In those instances, the administration of a subset of EF tasks may be preferable to not administering any at all. Quantifying the relative decrements in the maximal reliability of a latent variable of EF that resulted from using different subsets of tasks represents one method for identifying optimal short forms. The tendency for some three-task combinations to differ only slightly from the full set of tasks is indicative of the fact that some tasks are better indicators of the latent variable than are others (i.e., they have stronger factor loadings). Although our tasks were originally conceptualized as indicators of potentially distinct subcomponents of EF (i.e., inhibitory control, working memory, attention shifting), there is no consistent empirical evidence to support the differentiation of these

subcomponents in early childhood. Nonetheless, researchers may still prefer to select tasks based on the facet of EF that each was intended to measure (e.g., whereas some researchers may choose 3 tasks that all purportedly measure inhibitory control, others may choose one task each for inhibitory control, working memory, and attention shifting). Substantive preference may inform the selection among 3-task combinations with comparable levels of maximal reliability.

This study was characterized by at least four limitations. First, our tasks do a relatively better job of measuring EF ability in children characterized by low to average but not high levels of true ability (Willoughby, Wirth, & Blair, 2011; Willoughby, Wirth, et al., 2012). This may have contributed to our estimates of maximal reliability. Second, the EF tasks were administered in a fixed order at each assessment (a necessity of the parent study). Counter-balancing the task order would have been a preferable design for the present purposes. Third, our EF assessment was embedded in home visits that included a large number of parent and child activities and that occurred at various points in the day. Reducing visit length and standardizing the times of the day that tasks were completed would have been more desirable for the present purposes. To the extent that test fatigue contributed to child performance on EF tasks, future studies might also consider randomly assigning children to complete subsets of tasks and focus on between-group comparisons to inform maximal reliability. Fourth, the estimates of maximal reliability that have been reported here are sample specific. Although this study benefitted from a large sample size, the specific values obtained in future studies that use this battery, may differ—though we would expect that these values would fall within the 95% confidence intervals that we have reported.

Executive function specifically and self-regulation more generally are currently “hot” topics in early childhood development. The recent literature has emphasized the dimensionality of EF tasks, including whether (and when) the dimensionality of EF tasks may change during the transition from early to middle childhood. As consensus on the dimensionality of EF is reached, it will be important that researchers not confuse the presence of good fitting CFA models as evidence that the latent variable of EF has been measured well. To the extent that the maximal reliability of EF constructs is poor, this undermines efforts to describe the developmental course of EF, to validate the construct of EF, and the use of EF as an outcome for intervention or training studies. In addition, a focus on changes in maximal reliability represents a principled method for the derivation of short forms of extant batteries. This solves a practical problem routinely encountered in research and clinical settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support for this research was provided by the National Institute of Child Health and Human Development grants R01 HD51502 and P01 HD39667, with co-funding from the National Institute on Drug Abuse. The Family Life Project Investigators include: Lynne Vernon-Feagans, University of North Carolina Chapel Hill (UNC-CH);

Martha Cox, UNC-CH; Clancy Blair, New York University; Peg Burchinal, UNC-CH; Linda Burton, Duke University; Keith Crnic, Arizona State University; Ann Crouter, PSU; Patricia Garrett-Peters, UNC-CH; Mark Greenberg, PSU; Stephanie Lanza, PSU; Roger Mills-Koonce, UNC-CH; and Michael Willoughby, UNC-CH.

Drs. Blair and Willoughby retain a copyright on all of the EF tasks described in this manuscript.

References

- Anderson, P.J. Towards a developmental model of executive function.. In: Anderson, V.; Jacobs, R.; Anderson, P.J., editors. *Executive functions and the frontal lobes: a lifespan perspective*. Psychology Press; New York: 2008. p. 3-22.
- Arnsten AFT. Catecholamine modulation of prefrontal cortical cognitive function. *Trends in Cognitive Sciences*. 1998; 2:436–447. [PubMed: 21227275]
- Arnsten AFT. Stress impairs prefrontal cortical function in rats and monkeys: role of dopamine D1 and norepinephrine alpha-1 receptor mechanisms. *Cognition, Emotion and Autonomic Responses: The Integrative Role of the Prefrontal Cortex and Limbic Structures*. 2000; 126:183–192.
- Bryck RL, Fisher PA. Training the Brain Practical Applications of Neural Plasticity From the Intersection of Cognitive Neuroscience, Developmental Psychology, and Prevention Science. *American Psychologist*. 2012; 67(2):87–100. doi: Doi 10.1037/A0024657. [PubMed: 21787037]
- Cragg L, Nation K. Self-ordered pointing as a test of working memory in typically developing children. *Memory*. 2007; 15(5):526–535. [PubMed: 17613795]
- Durston S, Thomas KM, Yang YH, Uluğ AM, Zimmerman RD, Casey BJ. A neural basis for the development of inhibitory control. *Developmental Science*. 2002; 5(4):F9–F16.
- Ettenhofer ML, Hambrick DZ, Abeles N. Reliability and stability of executive functioning in older adults. *Neuropsychology*. 2006; 20(5):607–613. [PubMed: 16938023]
- Garon N, Bryson SE, Smith IM. Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*. 2008; 134(1):31–60. [PubMed: 18193994]
- Gerardi Caulton G. Sensitivity to spatial conflict and the development of self-regulation in children 24-36 months of age. *Developmental Science*. 2000; 3(4):397–404.
- Gerstadt C, Hong Y, Diamond A. The relationship between cognition and action: performance of children 3 1/2 -7 years old on a Stroop-like day-night test. *Cognition*. 1994; 53:129–153. [PubMed: 7805351]
- Graham JW, Taylor BJ, Olchowski AE, Cumsille PE. Planned missing data designs in psychological research. *Psychological Methods*. 2006; 11(4):323–343. [PubMed: 17154750]
- Hancock, GR.; Mueller, RO. Rethinking construct reliability within latent variable systems.. In: Cudeck, R.; du Toit, S.; Sörbom, D., editors. *Structural Equation Modeling: Present and Future — A Festschrift in honor of Karl Jöreskog*. Scientific Software International, Inc.; Lincolnwood, IL: 2001.
- Hancock GR, Mueller RO. The Reliability Paradox in Assessing Structural Relations Within Covariance Structure Models. *Educational and Psychological Measurement*. 2011; 71(2):306–324. doi: Doi 10.1177/0013164410384856.
- Hughes C, Ensor R, Wilson A, Graham A. Tracking Executive Function Across the Transition to School: A Latent Variable Approach. *Developmental Neuropsychology*. 2010; 35(1):20–36. [PubMed: 20390590]
- Jacques S, Zelazo PD. The Flexible Item Selection Task (FIST): A measure of executive function in preschoolers. *Developmental Neuropsychology*. 2001; 20(3):573–591. [PubMed: 12002094]
- Kane MJ, Engle RW. Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology-General*. 2003; 132(1):47–70. [PubMed: 12656297]
- Miyake A, Friedman NP. The nature and organization of individual differences in executive functions: four general conclusions. *Current Directions in Psychological Science*. 2012; 21:8–14. [PubMed: 22773897]
- Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*. 2000; 41(1):49–100. [PubMed: 10945922]

- Nunnally, JC.; Bernstein, IH. *Psychometric Theory*. 3rd ed.. McGraw-Hill; New York: 1994.
- Penev S, Raykov T. On the relationship between maximal reliability and maximal validity of linear composites. *Multivariate Behavioral Research*. 2006; 41(2):105–126.
- Petrides M, Milner B. Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*. 1982; 20(3):249–262. [PubMed: 7121793]
- Rabbitt, P. Introduction: Methodologies and models in the study of executive function.. In: Rabbitt, P., editor. *Methodology of Frontal and Executive Function*. Psychology Press; 1997. p. 1-38.
- Raykov T. Estimation of maximum reliability: A note on a covariance structure modeling approach. *British Journal of Mathematical & Statistical Psychology*. 2004; 57:21–27. [PubMed: 15171799]
- Raykov T, Hancock GR. Examining change in maximal reliability for multiple-component measuring instruments. *British Journal of Mathematical & Statistical Psychology*. 2005; 58:65–82. doi: Doi 10.1348/000711005x38753. [PubMed: 15969840]
- Romine CB, Reynolds CR. Model of the development of frontal lobe functioning: Findings from a meta-analysis. *Applied Neuropsychology*. 2005; 12(4):190–201. [PubMed: 16422660]
- Schoemaker K, Bunte T, Wiebe SA, Espy KA, Dekovic M, Matthys W. Executive function deficits in preschool children with ADHD and DBD. *Journal of Child Psychology and Psychiatry*. 2012; 53(2):111–119. doi: DOI 10.1111/j.1469-7610.2011.02468.x. [PubMed: 22022931]
- Shing YL, Lindenberger U, Diamond A, Li SC, Davidson MC. Memory Maintenance and Inhibitory Control Differentiate from Early Childhood to Adolescence. *Developmental Neuropsychology*. 2010; 35(6):679–697. [PubMed: 21038160]
- Vernon-Feagans L, Cox M, Investigators FLPK. *The Family Life Project: An Epidemiological and Developmental Study of Young Children Living in Poor Rural Communities Monographs of the Society for Research in Child Development*. (in press).
- Wiebe SA, Espy KA, Charak D. Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology*. 2008; 44(2):575–587. [PubMed: 18331145]
- Wiebe SA, Sheffield T, Nelson JM, Clark CAC, Chevalier N, Espy KA. The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology*. 2011; 108(3):436–452. doi: DOI 10.1016/j.jecp.2010.08.008. [PubMed: 20884004]
- Willoughby MT, Blair CB. Test-retest reliability of a new executive function battery for use in early childhood. *Child Neuropsychology*. 2011; 17(6):564–579. [PubMed: 21714751]
- Willoughby MT, Blair CB, Wirth RJ, Greenberg M, Investigators FLP. The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. *Psychological Assessment*. 2010; 22(2):306–317. [PubMed: 20528058]
- Willoughby MT, Blair CB, Wirth RJ, Greenberg M, Investigators FLP. The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment*. 2012; 24(1):226–239. [PubMed: 21966934]
- Willoughby MT, Wirth RJ, Blair CB. Contributions of modern measurement theory to measuring executive function in early childhood: an empirical demonstration. *Journal of Experimental Child Psychology*. 2011; 108(3):414–435. [PubMed: 20553690]
- Willoughby MT, Wirth RJ, Blair CB, Investigators FLP. Executive function in early childhood: Longitudinal measurement invariance and developmental change. *Psychological Assessment*. 2012; 24(2):418–431. [PubMed: 22023561]

Figure 1. Dot Plot of Maximal Reliability and 95% CIs for All Three-Task Combinations Using the Mixed Age Sample

The dashed line represents the maximal reliability of EF as defined by the full set of six tasks ($\rho = .72$).

Figure 2.
Combined Dot Plots of Maximal Reliability and 95% CIs for Three-Task Combinations at
Ages 3, 4, and 5 Years.