Psychological Assessment
2016, Vol. 28, No. 3, 319–330

1040-3590/16/$12.00 http://dx.doi.org/10.1037/pas0000152

Measuring Executive Function in Early Childhood: A Case for Formative Measurement

Michael T. Willoughby
University of North Carolina at Chapel Hill

Clancy B. Blair
New York University

The Family Life Project Investigators
University of North Carolina at Chapel Hill, The Pennsylvania State University, Duke University, and The Arizona State University

This study tested whether individual executive function (EF) tasks were better characterized as formative or reflective indicators of the latent construct of EF. EF data that were collected as part of the Family Life Project (FLP), a prospective longitudinal study of families who were recruited at the birth of a new child (N = 1,292), when children were 3, 4, and 5 years old. Vanishing tetrad tests were used to test the relative fit of models in which EF tasks were used as either formative or reflective indicators of the latent construct of EF in the prediction of intellectual ability (at Age 3), attention-deficit hyperactivity disorder symptoms (at Ages 3 to 5 years), and academic achievement (at kindergarten). Results consistently indicated that EF tasks were better represented as formative indicators of the latent construct of EF. Next, individual tasks were combined to form an overall measure of EF ability in ways generally consistent with formative (i.e., creating a composite mean score) and reflective (i.e., creating an EF factor score) measurement. The test–retest reliability and developmental trajectories of EF differed substantially, depending on which overall measure of EF ability was used. In general, the across-time stability of EF was markedly higher when represented as a factor score versus composite score. Results are discussed with respect to the ways in which the statistical representation of EF tasks can exert a large impact on inferences regarding the developmental causes, course, and consequences of EF.

Keywords: executive function, early childhood, formative measurement

Executive functions (EFs) refer to a set of cognitive abilities that are important for organizing information, for planning and problem solving, and for orchestrating thought and action in support of goal-directed behavior (Blair & Ursache, 2011). Hence, the general referent EF refers to a wide range of interrelated abilities that serve integrative functions. Scientific interest in EF has grown exponentially over the last 25 years. For example, a search of the term *executive function* in the Web of Science (which accesses the Science Citation Index Expanded, Social Sciences Citation Index, and the Arts & Humanities Citation Index databases) identified 18 studies from 1985 to 1990 that used "executive function" in the title or keywords, compared with 7,445 studies that did so from 2006 to 2010.

Current Conceptualizations of the Construct of EFs

Despite the surge of multidisciplinary interest in EF, numerous questions about how to best measure the construct remain unanswered. For example, despite the potential ease of use, parent-ratings of children's EF behaviors correlate very poorly

This article was published Online First June 29, 2015.

Michael T. Willoughby, FPG Child Development Institute, University of North Carolina at Chapel Hill; Clancy B. Blair, Department of Applied Psychology, New York University; and The Family Life Project Investigators, University of North Carolina at Chapel Hill, The Pennsylvania State University, Duke University, and The Arizona State University.

Michael T. Willoughby is now at RTI International, Research Triangle Park, North Carolina.

The National Institute of Child Health and Human Development Grants R01 HD51502 and P01 HD39667, with cofunding from the National Institute on Drug Abuse, supported data collection. The Institute of Educational Sciences Grant R324A120033 supported data analysis and writing. The views expressed in this article are those of the authors and they do not necessarily represent the opinions and positions of the Institute of Educational Sciences, the Department of Education, or the National Institute of Child Health and Human Development. The Family Life Project Phase I Key Investigators include Lynne Vernon-Feagans, University of North Carolina at Chapel Hill; Martha Cox, University of North Carolina at Chapel Hill; Clancy B. Blair, The Pennsylvania State University; Peg Burchinal, University of North Carolina; Linda Burton, Duke University; Keith Crnic, The Arizona State University; Ann Crouter, The Pennsylvania State University; Patricia Garrett-Peters, University of North Carolina at Chapel Hill; Mark Greenberg, The Pennsylvania State University; Stephanie Lanza, The Pennsylvania State University; Roger Mills-Koonce, University of North Carolina at Greensboro; Debra Skinner, University of North Carolina at Chapel Hill; Emily Werner, The Pennsylvania State University; and Michael T. Willoughby, RTI International, Research Triangle Park, North Carolina.

Correspondence concerning this article should be addressed to Michael T. Willoughby, RTI International, Hobbs #349, 3040 Cornwallis Road, Research Triangle Park, NC 27709. E-mail: mwilloughby@rti.org

with children's performance on EF assessments (median correlation of $r = .19$ across 20 studies; see Toplak, West, & Stanovich, 2013). More troubling is evidence that performance-based indicators of EF are typically poorly to modestly correlated, despite being administered at the same time, using the same method, in the same setting, by the same person.[1] As we recently reported, the weak to modest correlations among performance-based indicators of EF (mean $r = .30$ for associations between tasks intended to measure EF or one of its subdomains—e.g., inhibitory control) were evident in studies that varied substantially with respect to participant age (3 to 70+ years of age) and the specific tasks used (Willoughby, Holochwost, Blanton, & Blair, 2014). These results suggested that weak to modest correlations among performance-based indicators may be a characteristic of the construct of EF and were not indicative of measurement deficiencies for a particular set of tasks or for a particular age group (e.g., young children). Hence, disagreements between rated and performance-based indicators notwithstanding, even the agreement among multiple performance-based indicators of EF is troublesome.

In the absence of a narrowly defined consensus definition, EFs have been described using a variety of metaphors. For example, EFs were recently likened to the airport traffic control system (Center on the Developing Child at Harvard University, 2011) and as the conductor of an orchestra (Espy et al., in press). Although heuristically useful, these metaphors risk perpetuating the idea that the brain has a dedicated system (e.g., an EF module) that is regionally bound to the prefrontal cortex. This conceptual framing is consistent with the characterization of EF as a latent variable that "gives rise to" (accounts for) the covariation of individual performance across a set of performance-based EF tasks. Moreover, this perspective closely conforms to the assumptions of factor analytic techniques, which are routinely used to represent individual differences in EF on the basis of individual performance across a battery of tasks.

An alternative characterization of EF is that it represents a range of specific cognitive abilities that depend on multiple distributed networks and brain-wide connectivity "hubs" (Cole et al., 2013; Petersen & Posner, 2012). From this perspective, the prefrontal cortex is important because of the dense interconnections it shares with other parts of the brain. For example, in the case of inhibitory control, Munakata et al. (2011) emphasized that different prefrontal regions played unique roles for distinct types of inhibition on the basis of their differential patterns of connectivity with other regions of the brain. Similarly, Chrysikou, Weber, and Thompson-Schill (2014) emphasized that the prefrontal cortex exerted top-down influences on other aspects of cognition and served as a filtering mechanism to bias bottom-up sensory information in ways that facilitate optimal behavioral responses that were sensitive to context. The important point is that there is no EF system or module. Rather, EF may be better characterized as an emergent property of individuals. This conceptual framing is consistent with the characterization of EF as a latent variable that is defined by (rather than giving rise to) individual performance across a set of performance-based tasks. This perspective does not correspond well with the use of factor analytic techniques as a statistical approach for representing individual differences across a set of performance-based EF tasks.

The overarching objective of this study is to explicate these contrasting perspectives on the way in which EF is conceptualized specifically as it informs the statistical modeling of the latent construct of EF. To date, virtually all studies have implicitly treated children's performance on individual EF tasks as reflective indicators of the construct of EF through their use of exploratory and confirmatory factor analysis. Here, we introduce an alternative conceptualization of the latent construct of EF, which characterizes individual EF tasks as formative (not reflective) indicators of the latent construct of EF. We use a combination of statistical and pragmatic evidence in order to demonstrate the potential utility of conceptualizing EF tasks as formative indicators of the latent construct of EF.

Reflective Versus Formative Indicators of Latent Variables

Latent variables that are exclusively defined by reflective indicators are characterized by paths that emanate from the latent construct into manifest indicators (see the top panels of Figures 1, 2, and 3). In contrast, latent variables that are exclusively defined by formative indicators are characterized by paths that emanate from the manifest indicators into the latent construct (see the bottom panels of Figures 1 through 3). Although the distinction between reflective and formative measurement is not new (Blalock, 1974; Fornell & Bookstein, 1982; Heise, 1972), the merits and pitfalls of these contrasting perspectives continue to be actively debated among psychometricians (Bollen & Bauldry, 2011; Diamantopoulos, Riefler, & Roth, 2008; Edwards, 2011; Howell, Breivik, & Wilcox, 2007b).

Three linked sets of ideas help to provide an intuitive understanding of the differences between latent constructs that are composed of reflective or formative indicators. First, latent variables that are represented using exclusively reflective indicators are characterized by that variation that is *shared* among those indicators. In contrast, latent variables that are represented using exclusively formative indicators are characterized by the *total* variation across those indicators. Second, whereas reflective constructs assume that indicators are positively correlated (and preferably of moderate to large magnitude), formative constructs make no assumptions about either the direction or magnitude of correlations between indicators. By extension, whereas traditional indices of the reliability are relevant for reflective constructs, they are irrelevant for formative constructs (Bollen, 1984; Bollen & Lennox, 1991). Third, reflective indicators of a latent construct are considered interchangeable; hence, the addition or removal of any indicator does not change the substantive meaning of the construct. In contrast, formative indicators are intended to represent multiple facets of the construct; hence, the addition or removal of any indicator has the potential to change the substantive meaning of the construct.

[1] Given our focus on the early childhood period, in which the preponderance of the current evidence indicates that EF is an undifferentiated (unidimensional) construct, we use the generic referent EF throughout. However, all of our arguments equally apply to the study of more narrowly defined subdimensions of EF—including inhibitory control (IC), working memory (WM), or attention shifting (AS)—that are more typically studied in older children and adults.

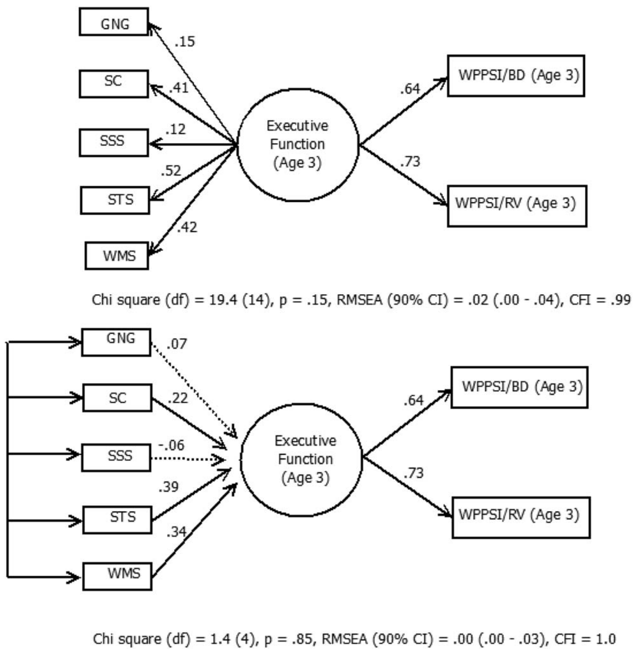

Figure 1. Reflective (top) and formative (bottom) indicators of executive function predicting Wechsler Preschool and Primary Scales of Intelligence subtests.

Differences between latent constructs that consist of (entirely) formative or reflective indicators can also be discerned through their equations. Following the notation of Bollen and Bauldry (2011), the equations for a latent construct with three reflective (i.e., "effect") indicators are

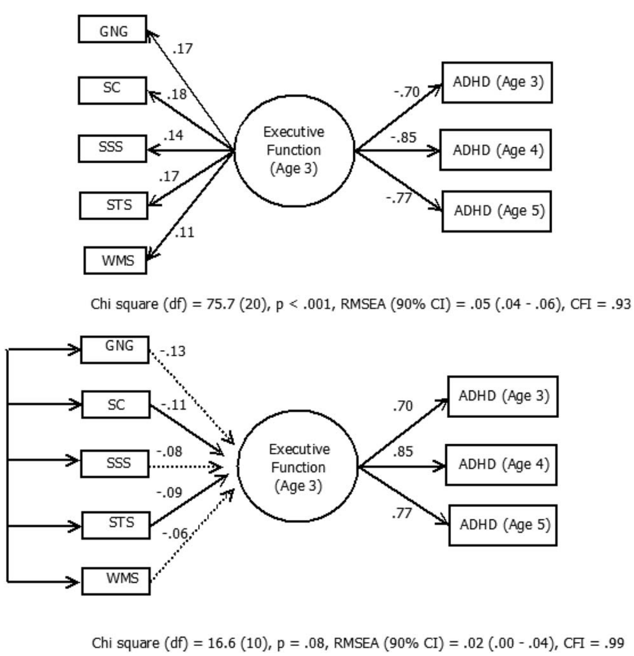

Figure 2. Reflective (top) and formative (bottom) indicators of executive function predicting attention-deficit hyperactivity disorder behaviors.

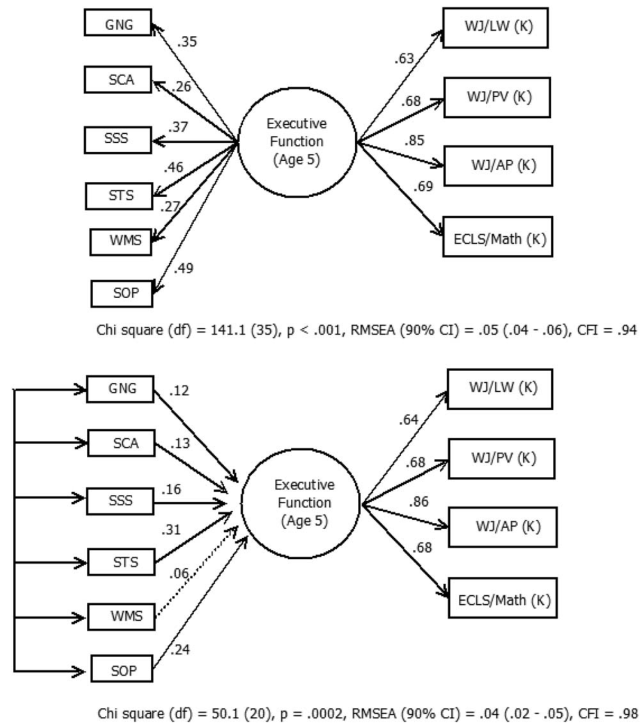

Figure 3. Reflective (top) and formative (bottom) indicators of executive function predicting academic achievement.

$$y_{1i} = \alpha_1 + \lambda_{11}\eta_{1i} + \varepsilon_{1i} \tag{1}$$

$$y_{2i} = \alpha_2 + \lambda_{21}\eta_{1i} + \varepsilon_{2i} \tag{2}$$

$$y_{3i} = \alpha_3 + \lambda_{31}\eta_{1i} + \varepsilon_{3i} \tag{3}$$

where y_{pi} is the *p*th indicator that depends on the latent construct, η_{1i}. The factors loadings, λ_p, represent structural coefficients that describe the magnitude of the association between each the latent construct and its indicators. The residual variances, ε_{pi}, reflect that part of the manifest indicator y that is not accounted for by the latent construct. Latent variables that are composed entirely of reflective indicators have as many equations as indicators. Moreover, reflective indicators are chosen to represent the theoretical definition of the latent construct of interest (i.e., they have conceptual unity; see Bollen & Bauldry, 2011). For comparison purposes, the equation for a latent construct with three formative (i.e., "causal") indicators is

$$\eta_{1i} = \alpha_\eta + \gamma_{11}x_{1i} + \gamma_{21}x_{2i} + \gamma_{31}x_{3i} + \zeta_{1i} \tag{4}$$

where x_{pi} is the *p*th indicator of the latent construct η_{1i}. The single residual variance, ζ, represents all of the influences of the latent construct, η_{1i}, that are not captured by the formative indicators. Latent variables that are composed of entirely formative indicators have a single equation with as many predictors as indicators. Like reflective indicators, formative indicators are expected to have conceptual unity. Bollen and Bauldry (2011) drew a further distinction between formative (causal) and so-called "composite" indicators. The equation for a three indicator composite construct is

$$C_{1i} = w_{10} + w_{11}x_{1i} + w_{12}x_{2i} + w_{13}x_{3i} \quad (5)$$

where x_{pi} is the *p*th indicator of the composite construct C_{1i}.The primary difference between composite variables (Equation 5) and latent variables that are defined entirely by formative indicators (Equation 4) is that composites do not include a disturbance term. That is, composites are exact linear combinations of their indicators. Moreover, there is no assumption that composite indicators necessarily have conceptual unity.

A third way to understand the differences between latent variables that consist of (entirely) formative (including causal and composite) and reflective (effect) indicators is with reference to their implied statistical representation. A latent construct that consists of entirely reflective indicators is represented using exploratory and confirmatory factor analytic models. A latent construct that consists of entirely formative indicators is represented using multiple indicator, multiple outcome models. A corollary point is that latent constructs that entirely consist of formative indicators are statistically underidentified and can only be estimated if two or more outcomes are available (MacCallum & Browne, 1993). This has generated debate regarding the inherent meaning of such latent constructs, which is beyond the scope of this article (see Bollen, 2007; Howell, Breivik, & Wilcox, 2007a; Howell et al., 2007b). Composite constructs are best represented using principle components analysis or using a simple aggregation (e.g., mean) of scores, which is analogous to a principle components analysis approach to scoring that applies unit weights.

In addition to practical and statistical differences, latent constructs that consist entirely of reflective and formative indicators may be understood to invoke different philosophies of science. Following Borsboom, Mellenbergh, and van Heerden (2003), latent constructs that are composed of reflective indicators imply a realist philosophical view in which latent variables are presumed to exist apart from and precede the measurement of indicator variables. In contrast, latent constructs that are composed of formative indicators may imply a constructivist philosophical view in which latent variables do not exist apart from observed measures, but instead reflect a summary of such measures.

Strategies for Differentiating Formative From Reflective Indicators

Three general approaches can be used to help determine whether EF is best construed as a formative or reflective latent variable. The first approach relies on the application of a series of decision rules (see, e.g., Coltman, Devinney, Midgley, & Venaik, 2008; MacKenzie, Podsakoff, & Jarvis, 2005). Theoretically, the essential questions ask (a) whether the latent construct is assumed to exist independent of the measures used or is solely a combination of indicators, (b) the direction of causality between indicators and the latent construct, and (c) whether a set of indicators "share a theme," are interchangeable, and whether the conceptual domain of construct changes based on the addition or omission of items. Empirically, the essential questions ask (a) about the magnitude of correlations among indicators, (b) the extent to which indicators share the same antecedents and consequences as the construct, and (c) what the best representation of indicators as formative or reflective indicators is. We have considered these questions elsewhere (Willoughby et al., 2014). Ultimately, the reliance on this narrative approach does not facilitate unambiguous inferences regarding whether a set of performance-based tasks are better characterized as formative or reflective indicators of the latent construct of EF.

Fortunately, there exists a statistical approach that can be used to formally test whether a latent construct is best characterized as exclusively formative, exclusively reflective, or some combination of indicators. The so-called vanishing tetrad test (VTT) has been developed by Bollen and colleagues (Bollen & Ting, 1993, 1998, 2000; Hipp, Bauer, & Bollen, 2005). Although a full description of this approach is beyond the scope of this article, the key idea is that although models that differ with respect to their type of indicator (formative, reflective) are not nested in the conventional sense (i.e., there is no set of parameter constraints that result in a latent variable that is defined by formative indicators to be subsumed by a latent variable that is defined by reflective indicators, or vice versa), they are often nested with respect to their vanishing tetrads. The VTT statistic can be used to evaluate the *global* fit for any SEM (Hipp et al., 2005; Hipp & Bollen, 2003), as well as to test the *relative* fit of competing models that are nested with respect to their tetrads, which is how it was used here (see Bollen, Lennox, & Dahly, 2009, for an extended example). The first objective of the proposed study was to reestimate variations of models that we have previously published in this journal (Willoughby, Blair, Wirth, & Greenberg, 2010, 2012) and to use nested VTTs to determine whether children's performance-based tasks were better characterized as a formative or reflective indicators of the latent construct of EF.

In addition to statistical model comparisons, we also considered pragmatic evidence to help inform questions about the optimal way to represent children's performance across a battery of performance-based EF tasks. For example, if the nested VTTs indicated that EF tasks were better represented as formative versus reflective indicators of the construct of EF, a related question would be whether and how this would impact our practical understanding of EF. Once again, this was addressed through a reanalysis of results regarding the test–retest reliability and patterns of developmental change in our battery of EF tasks, which had previously assumed that individual EF tasks were reflective indicators of the latent construct of EF (Willoughby & Blair, 2011; Willoughby, Wirth, Blair, & Family Life Project Investigators, 2012). In our previous retest study, we reported modest retest correlations for individual tasks (*r*s ≈ .60), but an exceptionally high retest correlation for the latent variable estimate of ability (φ = .95), across the 2-week interval. In our longitudinal study, we reported exceptionally high correlations for the latent variable estimate of EF across 1- to 2-year intervals (φs = .86 to .91), which substantially exceeded the 1- to 2-year stabilities for individual tasks. Although we attributed those results to the merits of latent variable estimation, we have subsequently begun to question the meaning of 2-week *and* 2-year stabilities of this magnitude, including whether these results were an artifact of factoring tasks that were modestly correlated. The second goal of the current study was to examine whether and how the 2-week retest reliability and 2-year stability would change had EF been conceptualized as a formative latent construct.

In sum, the overarching objective of this study was to consider two competing ways of representing the latent construct of EF. A combination of statistical and pragmatic evidence was marshalled in order to help inform this decision. The pragmatic evidence, in particular, was intended to help inform questions about whether and how practical conclusions about the stability and change in EF abilities in early childhood may differ as a function of the ways in which individual EF task scores were combined.

Method

Participants

The Family Life Project (FLP) was designed to study young children and their families who lived in two (Eastern North Carolina, Central Pennsylvania) of the four major geographical areas of the United States with high poverty rates (Dill, 2001). The FLP adopted a developmental epidemiological design in which sampling procedures were employed to recruit a representative sample of 1,292 children whose families resided in one of the six counties at the time of the child's birth. Low-income families in both states and African American families in North Carolina were oversampled (African American families were not oversampled in Pennsylvania because the target communities were at least 95% non-African American). Full details of the sampling procedure appear elsewhere (Vernon-Feagans, Cox, & Family Life Key Investigators, 2013).

Of those families interested and eligible and selected to participate in the study, 1,292 families completed a home visit at 2 months of child age, at which point they were formally enrolled in the study. In total, 1,121 (87% of the total sample) children completed an EF assessment at the Age 3, 4, and/or 5 year assessments. This includes those children for whom an in-home visit was completed (i.e., families who had moved more than 200 miles from the study area completed measures by phone, which precluded direct assessments of children) and those children who were able to complete at least one EF task during at least one of the three (i.e., Age 3, 4, and 5 year) home visits. Children who did not participate in any of the 3-, 4-, or 5-year EF assessments ($n = 171$) did not differ from those who did ($n = 1{,}121$) with respect to child race (37% vs. 43% African American; $p = .15$), child gender (56% vs. 50% male; $p = .19$), state of residence (36% vs. 41% residing in Pennsylvania; respectively, $p = .26$), or being recruited in the low-income stratum (77% vs. 78% poor; $p = .75$).

Procedures

Data for this study were drawn from home visits that occurred when study children were 3 (two visits), 4 (one visit), and 5 (one visit) years old, as well as a school visit during the kindergarten year. Home visits consisted of a variety of parent and child tasks (e.g., cognitive testing, interviews, questionnaires, and interactions). School visits consisted of a variety of direct child assessments and classroom observations. In this study, we make use of children's achievement testing that was collected in the kindergarten (spring) assessment.

Measures

Executive function task descriptions. The EF battery consisted of seven tasks. Because we have already described these task in multiple articles this journal, we provide only abbreviated descriptions here.

Working memory span (WMS). This span-like task required children to perform the operation of naming and holding in mind two pieces of information simultaneously (i.e., the name of colors and animals in pictures of "houses") and to activate one of them (i.e., animal name) while overcoming interference occurring from the other (i.e., color name). Items were more difficult as the number of houses (each of which included a picture of a color and animal) increased.

Pick-the-picture (PTP) game. This is a self-ordered pointing task presented to children with a series of two, three, four, and six pictures in a set. Children were instructed to continue picking pictures within each set until each picture had "received a turn." This task requires working memory because children have to remember which pictures in each item set they have already touched (spatial location of pictures changes across trials and was uninformative). The PTP was too difficult for many 3-year-olds and was therefore only administered at the 4- and 5-year assessments.

Silly sounds stroop. This task presented children with pictures of cats and dogs and asked children to make the sound opposite of that which was associated with each picture (e.g., meow when showed picture of a dog). 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 task presented children with a response card that had a picture of a car and boat. Initially, all test stimuli (pictures of cars or boats identical to that on the response card) were subsequently presented in locations that were spatially compatible with their placement on the response card (e.g., pictures of cars always appeared above the car on the response card). Subsequently, test items required a contralateral response (e.g., children were to touch their picture of the car despite the fact that it appeared above the boat). This task required inhibitory control as children have to override the spatial location of test stimuli with reference to their response card. The SC was administered at the 3-year assessment.

Spatial conflict arrows (SCA). This task was identical in format to the SC task, with the exception that the response card consisted of two black dots ("buttons") and the test stimuli were arrows that pointed to the left or right. Children were instructed to touch the button to which the arrow pointed. Initially, all left (right) pointing arrows pointed to the (left) right, but subsequently they pointed in the opposite direction. The SCA was administered at the 4- and 5-year assessments.

Animal go/no-go. This is a standard go/no-go task in which children were instructed to click a button (which made an audible sound) every time they saw an animal (i.e., go trials), except when it was a pig (i.e., no-go trials). Varying numbers of go trials appeared 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 required inhibitory control.

Something's-the-same game. This task presented children with a pair of pictures for which a single dimension of similarity was noted (e.g., both pictures were the same color). Subsequently, a third picture was presented and children were asked to identify which of the first two pictures was similar to the new picture. This task required the child to shift his or her attention from the initial labeled to a new dimension of similarity (e.g., from color to size).

Executive function task scoring. As previously discussed (Willoughby, Wirth, et al., 2012), EF 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), whereas 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 was on a common developmental scale.

Intellectual aptitude and academic achievement task descriptions.

Wechsler preschool and primary scales of intelligence (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).

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

Early childhood longitudinal program kindergarten (ECLS-K) math assessment[2]. The ECLS-K direct math assessment was designed to measure conceptual knowledge, procedural knowledge, and problem solving within specific content strands using items drawn from commercial assessments with copyright permission, and other National Center for Educational Statistics (NCES) studies (e.g., National Assessment of Educational Progress). The math assessment involves a two-stage adaptive design; all children are asked a common set of "routing" items, and their performance on these items informs the difficulty level of the item set that is administered following the completion of routing items. This approach minimizes the potential for floor and ceiling effects. Item-response-theory methods were used to create math scores, using item parameters that were published in an NCES working paper that reported the psychometric properties of the ECLS-K assessments (Rock & Pollack, 2002).

Analytic Strategy

The first research question was addressed by estimating three pairs of structural equation models. Each pair of models regressed two or more outcomes on the latent construct of EF; the models differed in whether individual EF tasks (i.e., EAP scores) were represented as formative or reflective indicators of the latent construct of EF. Each pair of models was nested with respect to their model implied vanishing tetrads. We output the model implied covariance matrices for each pair of models, which were utilized in conjunction with a SAS macro that was made available by Hipp and colleagues (2005) in order to conduct nested VTTs. These results provided an empirical test of the relative fit of models that differed with respect to whether EF was a reflective or formative latent construct.

The second set of results involved the creation of a three pairs of summary scores, one pair per assessment period, which represented a child's overall ability level on the battery of EF tasks. The first summary score was a factor score estimate of a child's ability and represented EF as a reflective construct. The second summary score was a mean score estimate of a child's ability and represented EF as a formative (i.e., composite) construct. Both factor and mean scores utilized as many EF tasks as were available for a given child at a given assessment, and children's performance on each individual EF task was indicated by their EAP score, which was corrected for measurement error. We considered differences in the retest reliability and developmental course of factor and mean scores using descriptive statistics (e.g., Pearson correlations) and latent curve models (Bollen & Curran, 2006). These results provided a pragmatic basis for understanding whether and how differences in the method of combining EF task scores influenced substantive conclusions about stability and change in the latent construct of EF over time.

All descriptive statistics were computed using SAS version 9.3, and all structural equation (including latent curve) models were estimated using Mplus version 7.1 (Muthén & Muthén, 1998–2013). Structural equation models used robust full information maximum likelihood estimation and took the complex sampling design (oversampling by income and race; stratification) into account. The SAS macro made available by Hipp et al. (2005) was used to conduct nested VTTs.

Results

VTTs

The first research question involved direct comparisons of models in which individual EF task scores were used as either causal (formative) or effect (reflective) indicators of a latent construct of EF that predicted multiple indicators of child functioning.

Age 3 EF tasks predicting Age 3 IQ subtests. The first pair of models regressed children's performance on two indicators of intellectual ability (i.e., Block Design and Receptive Vocabulary subtests of the WPPSI) from the Age 3 assessment on the latent construct of EF at Age 3 (cf. Willoughby et al., 2010). As summarized in Figure 1, both models fit the data well and both indicated that the latent construct of EF was significantly predictive of the WPPSI (see Figure 1). Whereas all five EF tasks contributed, albeit weakly, to the definition of the latent construct of EF in the reflective (i.e., effect indicator) model, only three of the five individual EF tasks uniquely contributed to the definition of the latent construct of EF in the formative (i.e., causal indicator) model (see top and bottom panels of Figure 1, respectively). In both models, the latent construct of EF explained 42% and 54% of the observed variation in WPPSI Block Design and Receptive Vocabulary scores, respectively. The nested VTT was statistically significant, $\chi^2(10) = 19.9$, $p = .03$ (see Table 1); this indicated

[2] See http://nces.ed.gov/ecls/kinderassessments.asp.

Table 1
Vanishing Tetrad Test Comparisons of Formative Versus Reflective Indicator Models of EF

Model	Description	n	Reflective $\chi^2(df)$	Reflective prob	Formative $\chi^2(df)$	Formative prob	Comparison $\chi^2(df)$	Comparison prob
1	EF @ Age 3 → WPPSI @ Age 3	1,079	23.0 (14)	.06	3.1 (4)	.54	19.9 (10)	.03
2	EF @ Age 3 → ADHD @ Age 3, 4, 5	1,157	60.9 (20)	<.001	29.2 (10)	.002	31.7 (10)	.002
3	EF @ Age 5 → Achievement @ Age 5	1,086	81.1 (35)	<.001	56.3 (20)	<.001	24.8 (15)	.10

Note. All values are aggregated across 500 replications; the vanishing tetrad chi-square test statistics and associated probability values in the Reflective and Formative columns represent tests of the null hypothesis that all of the model implied vanishing tetrads are zero. The test statistic and associated probability value in the Comparison column represents a nested model comparison of reflective versus formative models; statistically significant chi-square tests provide empirical support for the model with fewer vanishing tetrads (i.e., the formative model). EF = executive function; prob = probability; *df* = degrees of freedom; WPPSI = Wechsler Preschool and Primary Scales of Intelligence; ADHD = attention-deficit hyperactivity disorder.

that the data were better explained by the formative model (i.e., the model with fewer vanishing tetrads). That is, the nested VTT indicated that the formative indicator specification (bottom panel of Figure 1) fit the data better than the reflective indicator specification (top panel of Figure 1).

Age 3 EF tasks predicting parent-rated attention-deficit hyperactivity disorder (ADHD) at Ages 3, 4, and 5. The second pair of models regressed parent-rated ADHD at Ages 3 to 5 on the latent construct of EF at Age 3 (cf. Willoughby et al., 2010). As summarized in Figure 2, both models fit the data reasonably well and both indicated that the latent construct of EF was significantly predictive of ADHD. Whereas all five EF tasks contributed, albeit weakly, to the definition of the latent construct of EF in the reflective model, only two of the five individual EF tasks uniquely contributed to the definition of the latent construct of EF in the formative model (see top and bottom panels of Figure 2, respectively). The latent construct of EF explained 49%, 73%, and 60% of the observed variation in parent-reported ADHD scores at Ages 3, 4, and 5, respectively. The nested VTT was statistically significant, $\chi^2(10) = 31.7$, $p = .002$, which indicated that individual EF tasks were better characterized as formative than reflective indicators of the latent construct of EF.

Age 5 EF tasks predicting academic achievement indicators in kindergarten. The third pair of models regressed performance on four academic achievement tests during kindergarten on the latent construct of EF at Age 5 (cf. Willoughby, Blair, et al., 2012). As summarized in Figure 3, both models fit the data reasonably well and both indicated that the latent construct of EF was significantly predictive of academic achievement in kindergarten. Whereas all six EF tasks contributed, albeit weakly, to the definition of the latent construct of EF in the reflective model, five of the six individual EF tasks uniquely contributed to the definition of the latent construct of EF in the formative model (see top and bottom panels of Figure 3, respectively). The latent construct of EF explained 41%, 46%, 75%, and 47% of the observed variation in children's performance on the WJ III Letter-Word, WJ III Picture Vocabulary, WJ III Applied Problems, and ECLS-K Math achievement tests, respectively. The nested VTT was not statistically significant, $\chi^2(15) = 24.8$, $p = .10$. Although this implied that individual EF tasks were equally well characterized as either formative or reflective indicators of the latent construct of EF, we noted that the median (vs. mean) p value for the nested VTT test across the 500 replication was .06. This result is more similar to the previous two outcomes than different.

Pragmatic Results: Descriptive Statistics

Next, we considered the descriptive statistics for two summary variables of overall EF performance—that is, factor score estimates and mean scores—at each age. The within- and across-time correlations between these alternative scoring methods appear in Table 2. Two points were noteworthy. First, although both factor and means scores appeared to exhibit linear change from Age 3 to 5 years, the across-time correlations for factor score estimates of EF ability (rs = .96 to .99) were substantially larger than those for mean score estimates of EF ability (rs = .32 to .59). The two scoring approaches provide divergent information regarding the across-time stability of the construct of EF. Second, despite pronounced differences in the across-time stability of factor and mean scores, the within-time correlations between factor and mean scores were relatively large, particularly at Ages 4 and 5 (rs = .67, .89, and .88 at Ages 3, 4, and 5 years, respectively). Within any assessment period, the two scoring approaches provide convergent information regarding individual differences in EF ability levels.

Pragmatic Results: Growth Curve Models

The most notable finding from Table 2 was the appreciably different across time correlations for factor versus mean score estimates of EF ability. In order to better characterize the apparent

Table 2
Descriptive Statistics for EF Battery Factor and Mean Scores at Ages 3, 4, and 5 Years

	1	2	3	4	5	6
1. FS (3)	—					
2. FS (4)	.99	—				
3. FS (5)	.96	.98	—			
4. MN (3)	.67	.56	.51	—		
5. MN (4)	.85	.89	.83	.37	—	
6. MN (5)	.75	.79	.88	.32	.59	—
n	973	1,009	1,036	973	1,009	1,036
Mean	−1.32	0.01	1.15	−0.54	−0.13	0.29
SD	0.26	0.85	0.82	0.54	0.51	0.48

Note. ns = 898 to 1,036; all ps < .001. EF = executive function; FS = factor score estimate of EF ability using all available tasks at a given assessment; MN = mean score estimate of EF ability using all available tasks at a given assessment; 3, 4, 5 = Age 3, 4, and 5 year assessments; *SD* = standard deviation.

differences in the stability and change of EF ability from Age 3 to 5 years, we estimated latent growth curve (LGC) models separately for factor and mean scores of EF. A linear LGC fit the mean scores extremely well, $\chi^2(1) = 1.2$, $p = .27$, root mean square error of approximation (RMSEA) = .01, 90% confidence interval (CI) [.00, .08], comparative fit index (CFI) = 1.0. The mean and variance of the intercept ($\mu_{Int} = -.05$, $p < .001$; $\varphi_{Int} = .12$, $p < .001$), which corresponded to the Age 4 assessment, and the linear slope ($\mu_{Slope} = .41$, $p < .001$; $\varphi_{Slope} = .04$, $p < .001$) were statistically significant. That is, there was significant variability in average ability at Age 4 and in the rate of linear change from Age 3 to 5 years. Individual differences in intercepts and slopes were also positively, albeit modestly, correlated, $\varphi_{Int,Slope} = .27$, $p = .002$; children with higher levels of EF ability (as indicated by mean scores across tasks) at Age 4 tended to have faster rates of linear growth in ability from Age 3 to 5 years. The residual variances for the mean scores were statistically significant at Ages 3 ($\varepsilon = .59$, $p < .001$) and 4 ($\varepsilon = .53$, $p < .001$), but not Age 5 ($\varepsilon = .07$, $p = .32$); the corresponding R^2s for mean scores were .42, .47, and .93 at Ages 3, 4 and 5, respectively.

When the identical parameterization was applied to the factor score estimates of overall EF ability, the LGC model fit poorly, $\chi^2(1) = 235.4$, $p < .001$, RMSEA = .45, 90% CI [.41, .51], CFI = .95, and the residual covariance matrix was nonpositive definite because of negative variance estimates for factor score indicators at Ages 3 ($\varepsilon = -.20$, $p < .001$) and 5 ($\varepsilon = -.58$, $p < .001$). The model was reestimated constraining these negative variance estimates to 0; however, model fit was still very poor, $\chi^2(3) = 2101.3$, $p < .001$, RMSEA = .79, 90% CI [.76, .82], CFI = .55. Given poor model fit, none of the parameter estimates were trustworthy; however, we noted that the latent correlation between intercepts and slopes approached unity, $\varphi_{Int,Slope} = .98$, $p < .001$, which was consistent with the large correlations reported in Table 2. In a final effort to obtain a model with acceptable fit, we reparameterized the LGC model by fixing the factor loadings to 0 and 1 at the Age 3 and 5 assessments and freely estimating the factor loading at the Age 4 year assessment. This parameterization permitted nonlinear change in means across time (Bollen & Curran, 2006), which we determined was optimal in our previous work that involved a second-order LGC (Willoughby, Wirth, et al., 2012). Although model fit was improved, it was still extremely poor, $\chi^2(2) = 1495.8$, $p < .001$, RMSEA = .82, 90% CI [.78, .85], CFI = .68. Once again, given poor model fit, none of the parameter estimates were trustworthy, though we again observed a latent correlation between intercepts and slopes that approached unity, $\varphi_{Int,Slope} = .92$, $p < .001$.

Pragmatic Results: Retest Reliability

We previously reported the results of a 2-week test–retest study of the EF battery involving 140 study participants at the Age 4 year assessment. In that study, we noted that whereas the 2-week retest reliability of individual tasks was modest (rs $\approx .60$), the correlation between latent variables representing ability across a 2-week retest period approached unity, $\varphi_{Retest} = .95$, $p < .001$ (Willoughby & Blair, 2011). Here, we report the 2-week retest correlation of the factor and mean score estimates of EF ability as rs of .99 and .76, respectively (both ps $< .001$). Following the method of Raghunathan, Rosenthal, and Rubin (1996), the retest correlation was stronger for factor than mean score estimates, $z = 39.2$, $p < .001$. Nonetheless, in both approaches, the aggregation of performance across the battery of tasks (as factor or mean scores) resulted in an improvement in retest reliability relative to when individual scores were considered alone. It is noteworthy that when EF task performance was summarized as factor scores, the *2-week* stability at the Age 4 year assessment was nearly identical to the *2-year* stability from Age 3 to 5 years (rs = .99 and .96, respectively). In contrast, when EF task performance is summarized using mean scores, the corresponding 2-week and 2-year stability estimates were both smaller and differ in magnitude (rs = .76 and .32, respectively).

Discussion

Although the benefits of modeling EF as a latent variable are well established, virtually all previous advice has advocated for the use of confirmatory factor analytic methods in which EF tasks are used as reflective indicators (Ettenhofer, Hambrick, & Abeles, 2006; Miyake et al., 2000; Wiebe, Espy, & Charak, 2008). The primary objective of this study was to investigate whether performance-based tasks may be better represented as formative indicators. Comparisons between three pairs of structural equation models, which considered children's intellectual function, academic achievement, and parent-rated ADHD behaviors as outcomes, consistently indicated that EF tasks were best represented as formative indicators. Descriptive results demonstrated how substantive conclusions regarding the retest reliability and the patterns of development change in EF in early childhood differed substantially depending on whether EF tasks are combined as mean (consistent with formative indicator) versus factor (consistent with reflective indicator) scores.

The initial motivation for considering the distinction between formative and reflective measurement of the latent construct of EF resulted from our observations of low to modest intercorrelations among children's performance on individual EF tasks in both our own and others' work (Willoughby et al., 2014). Previously, we observed that modest correlations between individual EF task scores were associated with modest levels of maximal reliability among the latent variable of EF (Willoughby, Pek, & Blair, 2013). Modest levels of maximal reliability indicate that the use of three to five EF tasks as indicators of a latent variable do a relatively poor job of representing (or "communicating") individual differences in the latent construct (Hancock & Mueller, 2001). By implication, modest levels of maximal reliability necessitate the administration of substantially more tasks (indicators) to measure a construct than has typically been the case or the development of new performance-based indicators that exhibit stronger intercorrelations. However, consideration of the magnitude of EF task intercorrelations, the focus on maximal reliability, and the suggestion that researchers should administer substantially more (or better) EF tasks in order to improve the maximal reliability of the latent construct of EF are all predicated on an implicit assumption of reflective measurement. To the extent that performance-based tasks are better construed as formative indicators of the latent construct of EF, all of these ideas are irrelevant. From the perspective of formative measurement, the magnitude of task intercorrelations is uninformative, maximal reliability is not a relevant metric for evaluating how well tasks represent individual difference in true ability level, and the administration of more tasks does not necessarily improve the quality of measurement.

Despite the substantial differences between formative and reflective perspectives of measurement, no methods exist that unequivocally delineate which perspective is correct; moreover, it is entirely conceivable that some constructs may be optimally represented using a combination of formative and reflective indicators. In the absence of a definitive strategy for distinguishing whether EF tasks are best conceptualized as formative versus reflective indicators, we considered conceptual, pragmatic, and statistical evidence. As noted at the outset, researchers have proposed a series of conceptual questions that may help inform whether a set of measures are better construed as causal or effect indicators of a particular construct. Conceptually, EF refers to a broad set of interdependent cognitive abilities that serve organizing and integrative functions. However, when performance-based tasks are modeled as reflective indicators, it is not clear that the resulting latent variable accurately represents its intended conceptual function. Rather than characterizing EF as the combination (summation) of a constituent set of skills, reflective indicator models represent EF more narrowly as that variation that is shared across a set of tasks. It is the mismatch between the conceptual definition of EF and the statistical representation of EF using reflective indicators that is the overarching concern of this study. We conjecture that formative indicator models provide a statistical representation of EF that is more compatible with the intended conceptual definition.

Empirical support for conceptualizing tasks as formative indicators of the construct of EF was evident from VTT of competing models. To be clear, although the VTTs provide an indication of whether a model that consists entirely of reflective indicators is consistent with the data (as evidenced by a nonsignificant VTT chi-square test statistic), a statistically significant VTT does not necessarily imply that (all of) the indicators are necessarily formative—though it is consistent with this as a possibility. A closer inspection of the results of VTTs that were used to compare models that represented EF as formative versus reflective indicators revealed a number of important points. First, both formative and reflective indicator models exhibited an acceptable fit to the observed data; hence, global model fit is not a criterion that can be used to determine which specification is preferred. Second, the regression coefficients linking the latent construct of EF to the outcomes (e.g., IQ subtests, ADHD, achievement tests) were identical irrespective of whether EF tasks were represented as formative or reflective indicators; hence, this is also not a criterion that can be used to determine which specification is preferred. Third, the formative and reflective indicator models differed in the model-implied covariance structure among the EF tasks. In the formative (causal indicator) specification, no constraints were made regarding the covariance structure of the individual EF tasks—all possible pairwise covariances were freely estimated. In the reflective (effect indicator) specification, the covariance structure among EF indicators is implied entirely through their shared association with a latent variable. If all possible pairwise covariances were introduced between the residual variances, the formative and reflective models would be chi square equivalent models (rendering VTTs useless). Fourth, for each of the three sets of outcomes that were considered, when EF tasks were specified as reflective indicators of the latent construct of EF, all of the tasks contributed to the definition of the construct (i.e., all of the factor loadings were statistically significant, albeit of modest magnitude). In contrast, when EF tasks were specified as formative indicators of the latent construct of EF, only a subset of the tasks contributed to the definition of the construct. The determination of which causal indicators are significant indicators of the latent construct of EF will depend on the outcomes being considered. Although this is a frequently noted limitation of formative models (Edwards, 2011; Howell et al., 2007b), it is not a perspective that is shared by everyone (Bollen, 2007; Bollen & Bauldry, 2011).

In light of evidence from the nested VTTs, we were interested in whether and how our previous substantive conclusions regarding the retest reliability and developmental change in EF would change from the perspective of formative and reflective measurement. To facilitate these comparisons, we compared results from models that approximated the latent variable of EF using either mean or factor scores across all available tasks at each assessment. A clear and divergent pattern of results were evident for these two scoring approaches. The factor score approach, which approximated reflective measurement, implied that the 2-week stability of EF was nearly perfect and that the 1 to 2 year stabilities of EF were approximately .90. Moreover, none of the estimated growth curve models provided an adequate fit to factor score estimates of EF ability across time, which constrains the types of future questions that can be asked of these data (e.g., predictors of individual differences in the level and rate of change in EF). These results implied that although EF develops (improves) between 3 and 5 years of age, individual differences in EF ability were (nearly) completely determined by Age 3 and were (nearly) completely preserved across repeated assessments that span intervals as short as 2 weeks and as long as 2 years. We conjecture that the extraordinarily high stability of EF factor scores across time was an artifact of factoring tasks that were weakly correlated. In contrast, the mean score approach, which approximated formative measurement, implied that the 2-week and 2-year stabilities (rs $= .76$ and .32, respectively) differed appreciably in magnitude, in a manner consistent with expectation (i.e., the longer the span of intervening time, the less correlated a construct should be, particularly if measured during a period of developmental change). Moreover, growth curve models fit the data well, with evidence for significant interindividual differences in both level and rates of change in EF across time.

Although we fully acknowledge that simple comparisons of these results do not provide a scientifically convincing approach for determining which scoring approach is most appropriate, we find the differences in results to be remarkable. Clearly, in our data (and perhaps other data), the decision about whether to use factor or mean scoring approaches for characterizing children's ability across a battery of EF tasks will fundamentally affect the inferences drawn about the nature, development, and malleability of EF in early childhood. Practically speaking, there is strong interest in identifying and developing strategies that enhance EF in children for the betterment of society (Diamond, 2012). The ability to detect effective strategies will be impacted by the ways in which EF is conceptualized, measured, and modeled. Pragmatically, we favor the mean scoring (formative perspective) approach because the results conform to expectations about the stability and change in EF that are consistent with the broader literature. Moreover, this approach facilitates our ability to ask questions about both the antecedents and consequences of trajectories of EF across time.

Study Limitations

This study was characterized by two limitations. First, we have presented the distinction formative and reflective latent constructs as a dichotomy; all EF tasks were conceptualized as either exclusively causal or effect indicators. However, it is entirely reasonable to represent latent variables as a mix of causal and effect indicators. We did not consider this possibility because we did not have a conceptually defensible rationale for considering some of our tasks as causal and others as effect indicators. Second, we contrasted inferences that resulted when EF tasks were represented as mean versus factor scores. In this case, mean and factor scores were intended to approximate formative and reflective measurement, respectively. However, as noted at the outset, the mean scoring approach is more accurately represented as a composite variable. Bollen and Bauldry (2011) make a clear distinction between composites and causal indicator latent constructs that we muddled here.

Challenges Associated With Formative Indicator Models

In the business (management, marketing) research literature, the full gamut of opinions on formative measurement is evident (Diamantopoulos, 2008; Diamantopoulos et al., 2008; Edwards, 2011). Because most readers will likely not be familiar with that literature, we briefly summarize four of the more vexing challenges of adopting a formative measurement perspective for combining individual EF tasks into an overall score. First, latent constructs that are composed entirely of formative (causal) indicators are not statistically identified; that is, irrespective of whether one assumes that EF tasks are best characterized as "causing" versus "being caused by" the latent construct of EF, latent variables are inestimable unless they have two effect indicators or, equivalently, two outcomes (MacCallum & Browne, 1993). This presents a practical problem, as the very nature of the latent construct of EF is nonconstant—it is always defined in part by the reflective indicators (or equivalently outcomes) being used to identify it. This problem can be circumvented by aggregating performance across individual EF tasks using mean scores (or equivalently principle components analysis), as we did here, but does so at the cost of making simplifying assumptions and leaving the latent variable framework (Bollen & Bauldry, 2011).

Second, formative constructs are sometimes criticized as "not measurement" (Edwards, 2011; Howell et al., 2007a, 2007b; Wilcox, Howell, & Breivik, 2008). Traditional metrics of internal consistency and maximal reliability are not applicable. Similarly, our recent reliance on maximal reliability estimates in order to create short forms of our EF task battery was predicated on the assumption that tasks were effect indicators of EF (Willoughby et al., 2013). To the extent that EF tasks are better construed as formative indicators of the construct of EF, the observed pattern of task correlations is uninformative for the creation of short forms of the battery (this is replaced by appealing to conceptual arguments about which facets of the construct are prioritized).

Third, in a related vein, formative constructs have been criticized because they often assume that task indictors are measured without error. This criticism can be made against the majority of applied research in the social and behavioral sciences that is based on sum or mean scores (e.g., any scoring approach that does not explicitly attend to measurement error). This was not a problem in our study, as our EF tasks that had already been purged of measurement error prior to their use here (Willoughby, Wirth, et al., 2012). More generally, by failing to attend to the measurement error of formative indicators, one risks creating formative (or composite) constructs that conflate true score variation with measurement error.

Fourth, in the context of reflective measurement, the establishment of longitudinal measurement invariance is a necessary precondition for modeling change across time (Widaman, Ferrer, & Conger, 2010); indeed, this was a focus of our earlier efforts that were published in this journal (Willoughby, Wirth, et al., 2012). To the extent that the measurement properties of a latent construct change across time, mean level changes are ambiguous. The extension of longitudinal measurement invariance to the case of formative constructs is less clear. Hypothetically, one could test for the plausibility of imposing across time constraints on the coefficients that relate formative indicators to the latent construct. However, these models are not estimable because of the underidentification problem that was described earlier. The only known work-around for this problem is to incorporate two or more reflective indicators into the formative construct and to test for longitudinal invariance of these reflective indicators prior to testing constraints regarding the contribution of formative indicators across time (Diamantopoulos & Papadopoulos, 2010). To be clear, although this approach was proposed for the situation involving cross-group comparisons, we are suggesting that it may generalize to longitudinal settings.

Conclusions

The recent proliferation of transdisciplinary research involving EF underscores the importance that has been attributed to this construct as an indicator of health and well-being. Nonetheless, a close reading of this literature suggests that this is an area in which the ideas are better than the measurement. Conceptual definitions of EF characterize it as a construct that subsumes a broad array of cognitive abilities that, collectively, facilitate engagement in novel problem solving efforts and enhance self-management. The primary objective of this study was to highlight an apparent lack of conformability between these conceptual definitions of EF and the use of psychometric approaches for combining EF task scores that assume reflective measurement. The combination of conceptual, pragmatic, and statistical evidence that was presented here suggests that performance-based measures may be better characterized as formative indicators of the latent construct of EF. Decisions about how to combine EF task scores will directly impact the types of inferences that will be made regarding the developmental origins, developmental course, and developmental outcomes of EF. Although we are unable to offer definitive conclusions, the intent of this study was to encourage other research groups that utilize performance-based indicators of EF to consider the distinction between formative and reflective measurement in their own work. More generally, our results point to the possibility that the construct of EF may not be well-suited to conventional measurement wisdom. Although this is neither an indictment of the construct of

EF nor of modern test theory, it is illustrative of problem that was first noted over two decades ago regarding the potential mismatch that can occur when the conceptualization of a psychological construct does not conform to the dominant statistical methods for representing it (Bollen & Lennox, 1991).

References

Blair, C. B., & Ursache, A. (2011). A bidirectional model of executive functions and self-regulation. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation* (2nd ed., pp. 300–320). New York, NY: Guilford Press.

Blalock, H. M. (Ed.). (1974). *Measurement in the social sciences: Theories and strategies*. Chicago, IL: Aldine.

Bollen, K. A. (1984). Multiple indicators: Internal consistency or no necessary relationship? *Quality & Quantity: International Journal of Methodology, 18,* 377–385. http://dx.doi.org/10.1007/BF00227593

Bollen, K. A. (2007). Interpretational confounding is due to misspecification, not to type of indicator: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods, 12,* 219–228. http://dx.doi.org/10.1037/1082-989X.12.2.219

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16,* 265–284. http://dx.doi.org/10.1037/a0024448

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314. http://dx.doi.org/10.1037/0033-2909.110.2.305

Bollen, K. A., Lennox, R. D., & Dahly, D. L. (2009). Practical application of the vanishing tetrad test for causal indicator measurement models: An example from health-related quality of life. *Statistics in Medicine, 28,* 1524–1536. http://dx.doi.org/10.1002/sim.3560

Bollen, K. A., & Ting, K. F. (1993). Confirmatory tetrad analysis. *Sociological Methodology, 23,* 147–175. http://dx.doi.org/10.2307/271009

Bollen, K. A., & Ting, K. F. (1998). Bootstrapping a test statistic for vanishing tetrads. *Sociological Methods & Research, 27,* 77–102. http://dx.doi.org/10.1177/0049124198027001002

Bollen, K. A., & Ting, K. F. (2000). A tetrad test for causal indicators. *Psychological Methods, 5,* 3–22. http://dx.doi.org/10.1037/1082-989X.5.1.3

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Center on the Developing Child at Harvard University. (2011). *Building the brain's "air traffic control" system: How early experiences shape the development of executive function: Working Paper No. 11*. Retrieved from http://www.developingchild.harvard.edu

Chrysikou, E. G., Weber, M. J., & Thompson-Schill, S. L. (2014). A matched filter hypothesis for cognitive control. *Neuropsychologia, 62,* 341–355.

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience, 16,* 1348–1355. http://dx.doi.org/10.1038/nn.3470

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research, 61,* 1250–1262. http://dx.doi.org/10.1016/j.jbusres.2008.01.013

Diamantopoulos, A. (2008). Formative indicators: Introduction to the special issue. *Journal of Business Research, 61,* 1201–1202. http://dx.doi.org/10.1016/j.jbusres.2008.01.008

Diamantopoulos, A., & Papadopoulos, N. (2010). Assessing the cross-national invariance of formative measures: Guidelines for international business researchers. *Journal of International Business Studies, 41,* 360–370. http://dx.doi.org/10.1057/jibs.2009.37

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61,* 1203–1218. http://dx.doi.org/10.1016/j.jbusres.2008.01.009

Diamond, A. (2012). Activities and programs that improve children's executive functions. *Current Directions in Psychological Science, 21,* 335–341. http://dx.doi.org/10.1177/0963721412453722

Dill, B. T. (2001). Rediscovering rural America. In J. R. Blau (Ed.), *Blackwell companions to sociology* (pp. 196–210). Malden, MA: Blackwell.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14,* 370–388. http://dx.doi.org/10.1177/1094428110378369

Espy, K., Clark, C., Chevalier, N., Nelson, J., Sheffield, T., Garza, J., . . . Wiebe, S. (in press). The changing nature of executive control in preschool. *Monographs of the Society for Research in Child Development*.

Ettenhofer, M. L., Hambrick, D. Z., & Abeles, N. (2006). Reliability and stability of executive functioning in older adults. *Neuropsychology, 20,* 607–613. http://dx.doi.org/10.1037/0894-4105.20.5.607

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research, 19,* 440–452. http://dx.doi.org/10.2307/3151718

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Heise, D. R. (1972). Employing nominal variables, induced variables, and block variables in path analyses. *Sociological Methods & Research, 1,* 147–173. http://dx.doi.org/10.1177/004912417200100201

Hipp, J. R., Bauer, D. J., & Bollen, K. A. (2005). Conducting tetrad tests of model fit and contrasts of tetrad-nested models: A new SAS macro. *Structural Equation Modeling, 12,* 76–93. http://dx.doi.org/10.1207/s15328007sem1201_4

Hipp, J. R., & Bollen, K. A. (2003). Model fit in structural equation models with censored, ordinal, and dichotomous variables: Testing vanishing tetrads. *Sociological Methodology, 33,* 267–305. http://dx.doi.org/10.1111/j.0081-1750.2003.00133.x

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007a). Is formative measurement really measurement? Reply to Bollen (2007) and Bagozzi (2007). *Psychological Methods, 12,* 238–245. http://dx.doi.org/10.1037/1082-989X.12.2.238

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007b). Reconsidering formative measurement. *Psychological Methods, 12,* 205–218. http://dx.doi.org/10.1037/1082-989X.12.2.205

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin, 114,* 533–541. http://dx.doi.org/10.1037/0033-2909.114.3.533

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology, 90,* 710–730. http://dx.doi.org/10.1037/0021-9010.90.4.710

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100. http://dx.doi.org/10.1006/cogp.1999.0734

Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., & O'Reilly, R. C. (2011). A unified framework for inhibitory control. *Trends in Cognitive Sciences, 15,* 453–459. http://dx.doi.org/10.1016/j.tics.2011.07.011

Muthén, L. K., & Muthén, B. O. (1998–2013). *Mplus user's guide* (7th ed.) Los Angeles, CA: Author.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience, 35,* 73–89. http://dx.doi.org/10.1146/annurev-neuro-062111-150525

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods, 1,* 178–183. http://dx.doi.org/10.1037/1082-989X.1.2.178

Rock, D. A., & Pollack, J. M. (2002). *Early childhood longitudinal study: Kindergarten class of 1998–99 (ECLS-K), Psychometric Report for kindergarten through first grade*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications*. San Diego, CA: Sattler Publisher, Inc.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry, 54,* 131–143. http://dx.doi.org/10.1111/jcpp.12001

Vernon-Feagans, L., Cox, M., & Family Life Key Investigators. (2013). 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, 78,* 1–150.

Wechsler, D. (2002). Wechsler Preschool and Primary Scale of Intelligence (WPPSI-III; 3rd ed.). San Antonio, TX: The Psychological Corporation.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4,* 10–18. http://dx.doi.org/10.1111/j.1750-8606.2009.00110.x

Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology, 44,* 575–587. http://dx.doi.org/10.1037/0012-1649.44.2.575

Wilcox, J. B., Howell, R. D., & Breivik, E. (2008). Questions about formative measurement. *Journal of Business Research, 61,* 1219–1228. http://dx.doi.org/10.1016/j.jbusres.2008.01.010

Willoughby, M., & Blair, C. (2011). Test–retest reliability of a new executive function battery for use in early childhood. *Child Neuropsychology, 17,* 564–579. http://dx.doi.org/10.1080/09297049.2011.554390

Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2010). The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. *Psychological Assessment, 22,* 306–317. http://dx.doi.org/10.1037/a0018708

Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2012). The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment, 24,* 226–239. http://dx.doi.org/10.1037/a0025361

Willoughby, M. T., Holochwost, S. J., Blanton, Z. E., & Blair, C. B. (2014). Executive functions: Formative versus reflective measurement. *Measurement: Interdisciplinary Research and Perspectives, 12,* 69–95. http://dx.doi.org/10.1080/15366367.2014.929453

Willoughby, M. T., Pek, J., & Blair, C. B. (2013). Measuring executive function in early childhood: A focus on maximal reliability and the derivation of short forms. *Psychological Assessment, 25,* 664–670. http://dx.doi.org/10.1037/a0031747

Willoughby, M. T., Wirth, R. J., Blair, C. B., & Family Life Project Investigators. (2012). Executive function in early childhood: Longitudinal measurement invariance and developmental change. *Psychological Assessment, 24,* 418–431. http://dx.doi.org/10.1037/a0025779

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Examiner's manual. Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Received April 21, 2014
Revision received March 13, 2015
Accepted April 3, 2015 ■