11-10-2017

Integrating Item Accuracy and Reaction Time to Improve the Measurement of Inhibitory Control Abilities in Early Childhood

Brooke E. Magnus
Marquette University, brooke.magnus@bc.edu

Michael T. Willoughby
RTI International

Clancy B. Blair
New York University

Laura J. Kuhn
University of North Carolina - Chapel Hill

Integrating Item Accuracy and Reaction Time to Improve the Measurement of Inhibitory Control Abilities in Early Childhood

Brooke E. Magnus
Marquette University, Milwaukee, WI
Michael T. Willoughby
RTI International, Research Triangle Park, NC
Clancy B. Blair
New York University, New York, NY
Laura J. Kuhn
Frank Porter Graham Child Development Institute Chapel Hill, NC

Abstract

Efforts to improve children's executive function are often hampered by the lack of measures that are optimized for use during the transition from preschool to elementary school. Whereas preschool-based measures often emphasize response accuracy, elementary school-based measures emphasize reaction time (RT)—especially for measures inhibitory control (IC) tasks that typically have a speeded

component. The primary objective of this study was to test in a preschool-aged sample whether the joint use of item-level accuracy and RT data resulted in improved scoring for three IC tasks relative to scores derived from accuracy data alone. Generally, the joint use of item-level accuracy and RT data resulted in modest improvements in the measurement precision of IC abilities. Moreover, the joint use of item-level accuracy and RT helped eliminate floor and ceiling effects that occurred when accuracy data were considered alone. Results are discussed with respect to the importance of scoring IC tasks in ways that are maximally informative for program evaluation and longitudinal modeling.

Keywords

executive function, inhibitory control, reaction time, psychometrics, task scoring

Executive functions (EFs) refer to a set of cognitive abilities that facilitate problem solving and goal-directed efforts. EF abilities develop across the life-course (De Luca et al., 2003; De Luca & Leventer, 2008), and early childhood represents an initial period of rapid developmental change. Improvements in EF abilities during early childhood are also understood to contribute to the successful transition to formal schooling (Shanmugan & Satterthwaite, 2016; Ursache, Blair, & Raver, 2012). Numerous efforts are underway to develop prevention and intervention activities that target young children, especially those from high-risk backgrounds who are at greatest risk for school failure (Blair & Diamond, 2008; Blair & Raver, 2015; Knudsen, Heckman, Cameron, & Shonkoff, 2006). However, one of the challenges of evaluating EF-focused interventions is the lack of performance-based measures of EF abilities that have been optimized for use during the transition from preschool to early elementary grades. Many EF tasks that work best for preschool-aged children become too easy for children during the transition to kindergarten (S. A. Carlson, 2005). This is especially true for inhibitory control (IC) tasks, for which floor and ceiling effects are well documented problems (Petersen, Hoyniak, McQuillan, Bates, & Staples, 2016).

IC is one of the three foundational components of EFs, the other two being cognitive flexibility and working memory (Diamond, 2013). Like EF, IC is a higher order construct that subsumes multiple component processes. Individual IC tasks have been distinguished in terms of whether they make delay versus conflict demands (S. M. Carlson & Moses, 2001), engage simple versus more complex processes (Garon, Smith, & Bryson, 2014), or involve tasks with varying degrees of motivational or affective (i.e., “hot” vs. “cool”) significance (Willoughby, Kupersmidt, Voegler-Lee, & Bryant, 2011). Notably, these distinctions are all special cases of Nigg’s (2000) taxonomy of eight types of IC. As summarized by Petersen et al. (2016), the most commonly used IC tasks in the childhood EF literature involve conflict (cf. delay), complex (cf. simple) demands, and are affectively cool (cf. hot). The tasks that we use in the current study are similarly characterized. Petersen et al. (2016) also noted that most IC tasks that are used with children are useful for less than 3 years, at which time they become insensitive to individual differences, often due to floor and ceiling effects. The limited time spans for which IC tasks are maximally useful undermines efforts to document normative developmental or intervention-related changes in IC abilities.

Petersen et al. (2016) proposed the use of multitask batteries and statistical procedures (i.e., formal tests of longitudinal measurement invariance) to build latent constructs of IC that can be used over longer spans of time. Although their approach has many merits, it makes strong conceptual and statistical assumptions about the nature of the higher order construct of IC (Willoughby, Holochwost,

Blanton, & Blair, 2014). Moreover, in many studies, time and resource constraints prohibit researchers from administering multiple IC tasks. In instances where only a few (perhaps even only one) IC tasks can be administered, an open question is whether IC tasks be scored differently to enhance their measurement precision and reduce the presence of floor and ceiling effects. One obvious approach involves the joint consideration of item accuracy and reaction time (RT) to inform task scoring. Whereas the accuracy of children's responses on IC is the primary source of information for scoring EF tasks that are used with preschool-aged children, the speed of children's responses (RT) is often the primary source of information for scoring EF tasks that are used with school-aged children.

The NIH Toolbox is notable in that it represents one of the first attempts to integrate accuracy and RT information into a single score for two EF tasks (i.e., for modified versions of the Dimensional Change Card Sort [DCCS] and Flanker tasks). This effort was motivated by a desire to have tasks that were suitable for use with children as young as 3 years old and that continue to yield useful task scores throughout adulthood (see Zelazo et al., 2013). For the DCCS task, participants are credited with a fixed point-value for every correctly answered item, yielding an accuracy score ranging from 0 to 5. Each participant's median RT across items is winsorized, log transformed, and converted to a RT score ranging from 0 to 5. Among participants who answer 80% or more of the items correctly, their task score is defined as the simple sum of their accuracy and RT scores (range: 0-10). For participants who do not complete 80% of items correctly (e.g., young children), their task score is defined solely by task accuracy (range: 0-5). An analogous procedure is used for the Flanker task. Although well-intentioned, the method of combining accuracy and RT data represents an ad hoc approach that is based on numerous untested assumptions (e.g., all accuracy items are assumed equally difficult; accuracy and RT scores are assumed to be equally important indicators of ability) The primary objective of the current study was to consider a model-based approach for integrating accuracy and RT information for three IC tasks.

Numerous psychometric models have been developed for the joint analysis of accuracy and RT data, and an exhaustive review is beyond the scope of this article (but see van der Linden, 2009, 2016, for a thorough review of RT modeling approaches within an item response theory framework). Notably, only a subset of all models that involve the joint analysis of accuracy and RT data have sought to integrate these sources of information into a single task score. Molenaar, Tuerlinckx, and van der Maas (2015) proposed a bivariate generalized linear item response theory (B-GLIRT) modeling approach that subsumes many of the models that integrate item-level accuracy and RT data into a single task score. The model considered here is a special case of the general family of B-GLIRT models.

The model that is the focus of this study is depicted in Figure 1. It involves a simultaneous analysis of item-level accuracy and RT data, which load onto two orthogonal latent factors—one of which represents general IC ability (as indexed by a particular task) and one of which represents a processing speed factor that is unrelated to IC. Whereas both the accuracy and RT items load on the ability factor, only the RT items load on the speed factor. This model parameterization, which can be referred to as "bifactor-like" (see Cai, Yang, & Hansen, 2011), partitions the observed variation in item-level RT into that which may be informative of IC ability versus that which reflects general speed of responding that is unrelated to IC ability (unexplained variation is measurement error). The statistical significance and magnitude of the factor loadings that relate item-level RT items to the IC ability factor provide a formal test of whether and to what extent item-level RT data improve the precision of measurement of IC beyond that obtained from an exclusive reliance on accuracy data. Similarly, the relative contribution of RT items to ability versus speed factors inform questions about the extent to which RT is primarily indicative of ability versus general speed of processing. This is important given that RT is often used as

the sole measure of IC ability in school-aged children, whose accuracy of responding is uniformly high. By demonstrating a general modeling approach for the joint analysis of item-level accuracy and RT data from IC tasks, we open the possibility for future studies to be able to use a common set of tasks for children as they transition from preschool to early elementary grades. This would have important implications for program evaluation efforts.

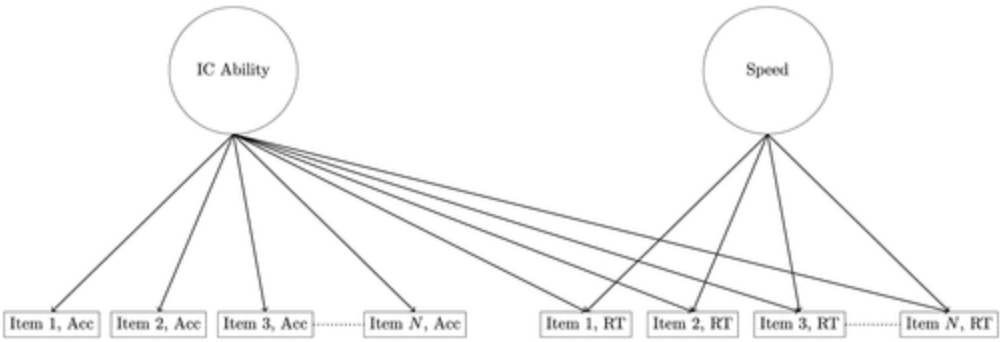

Figure 1. Path model for bivariate modeling of item-level RT and accuracy response data.

Note. RT = reaction time; IC = inhibitory control; Acc = accuracy.

Given the additional information that RTs may be able to provide about a child's level of IC ability, we hypothesized that the joint analysis of item-level accuracy and RT would result in improved measurement precision of IC ability (i.e., smaller standard errors [*SE*s] of factor scores) compared with the analysis of accuracy data alone. While the development of the proposed psychometric model was motivated by the hypothesis that RT information would enhance measurement precision, the current study is largely exploratory: Within the variance partitioning framework of the bifactor-like model, what can be learned from the joint analysis of item-level accuracy and RT?

Method

Participants and Procedure

This study involved a convenience sample of 844 preschool-aged children who lived in New York or North Carolina and who were recruited using a quota-based sampling procedure in order to ensure diversity with respect to race, ethnicity, household poverty, age, and gender. For full details on recruitment and study enrollment, see Camerota, Willoughby, Kuhn, and Blair (2016).

Due to the large number of tasks, a planned missing design was used to determine which subset of the IC tasks were assigned to each child. A total of 844 children completed at least one of the three tasks that were the focus of this study, and this constituted the analysis sample. Consistent with the quota sampling approach, participants were diverse with respect race (60% Caucasian, 31% African American, 7% Asian American, 1% Native American, and 1% Pacific Islander), ethnicity (20% Hispanic), gender (50% female), age (*M* = 4.4, *SD* = 0.70, range = 3.00-5.99 years), and household poverty level (25%, 21%, and 54% of children came from poor, near poor, and not poor households, respectively; poor, near poor, and not poor designations referred to total household income consistent with 0% to 100%, 100% to 200%, and >200% of the U.S. federal poverty threshold for a given household size).

Measures

Executive Function Touch (EF Touch)

EF Touch is a computerized battery of EF tasks that were initially created, administered, and extensively studied in paper-and-pencil (i.e., "flip book") formats (Willoughby, Blair, & Family Life Project Investigators, 2016; Willoughby, Blair, Wirth, Greenberg, & Family Life Project Investigators, 2010, 2012; Willoughby, Wirth, Blair, & Family Life Project Investigators, 2012). A history of the development of these tasks, including the rationale for and benefits of computerization is provided elsewhere (Willoughby & Blair, 2016). The *EF Touch* program runs in a Windows OS environment and requires two monitors. One standard monitor displays a script to the interviewer. A capacitive touch screen monitor records child responses (15 in. Planar capacitive touch monitors were used here). The battery is modular in nature (i.e., any number of tasks can be administered in any desired order), and each EF task takes 3 to 7 minutes to complete. This study focused on three IC tasks for the battery because these were the only tasks for which individual RT was collected (the other tasks did not have a speeded component).

Spatial Conflict Arrows

This IC task consisted of 36 items. Two "buttons" appear on the left and rightmost sides of the touch screen monitor. Children were instructed to touch the button to which the arrow is pointing. Three blocks of 12 arrows were depicted in which arrows either appeared above the button to which they are pointing (congruent condition), above the opposite button to which they are pointing (incongruent condition), or in mixed locations such that both congruent and incongruent trials appear in the same block. For the congruent items, the arrows are depicted laterally, building a prepotency to touch the button based on the location of the arrow. For the incongruent items, the arrows are depicted contralaterally. The spatial location of the arrow is no longer informative, and IC is required to override the previously established association. For this reason, only the 17 incongruent items were used to index task performance. Each item was presented for 4,000 milliseconds irrespective of the speed of a child's response. The accuracy of each response was recorded, as was the associated RT for each item (RT responses less than 300 milliseconds were deemed too fast to be valid and set to missing). On average, children completed 60% of these 17 incongruent items correctly (N = 644, M = 0.60, SD = 0.29, range = 0.00-1.00). For each item that a child answered correctly, the RT was recorded; RTs for incorrect responses were structurally missing. The two rightmost columns of Table 1 show the item-level percent correct and average RT for correct responses, respectively.

Table 1. Parameter Estimates and *SE*s for the Three IC Tasks.

	Standardized factor loadings (*SE*)			Intercepts/thresholds (*SE*)		Descriptive statistics	
Item #	Accuracy on IC ability	RT on IC ability	RT on speed	Accuracy threshold	RT intercept	Accuracy, % correct (*SD*)	RT (ms), *M* (*SD*)
Spatial Conflict Arrows							
1	0.62 (0.04)	−0.01 (0.06)	−0.47 (0.07)	−0.86 (0.12)	7.51 (0.03)	0.65 (0.48)	2026.79 (825.21)
2	0.67 (0.04)	0.00 (0.09)	−0.59 (0.07)	−0.84 (0.14)	7.50 (0.03)	0.64 (0.48)	1946.14 (689.57)
3	0.71 (0.04)	0.02 (0.07)	−0.58 (0.07)	−0.76 (0.16)	7.45 (0.03)	0.62 (0.49)	1868.55 (695.79)
4	0.72 (0.04)	−0.06 (0.08)	−0.58 (0.08)	−0.70 (0.15)	7.42 (0.03)	0.61 (0.49)	1791.67 (703.38)
5	0.32 (0.05)	0.30 (0.05)	−0.19 (0.08)	−0.24 (0.08)	7.55 (0.03)	0.56 (0.50)	2183.07 (858.32)
6	0.69 (0.04)	0.01 (0.08)	−0.46 (0.08)	−0.61 (0.13)	7.46 (0.03)	0.60 (0.49)	1841.01 (634.06)
7	0.76 (0.03)	0.01 (0.08)	−0.65 (0.05)	−0.83 (0.17)	7.44 (0.03)	0.62 (0.49)	1827.97 (708.89)
8	0.71 (0.03)	−0.18 (0.08)	−0.46 (0.08)	−0.62 (0.15)	7.49 (0.03)	0.60 (0.49)	1854.66 (726.25)
9	0.48 (0.04)	0.25 (0.06)	−0.17 (0.09)	−0.06 (0.10)	7.53 (0.04)	0.51 (0.50)	2185.43 (850.40)
10	0.69 (0.04)	−0.18 (0.08)	−0.47 (0.08)	−0.60 (0.14)	7.45 (0.03)	0.60 (0.49)	1789.05 (699.85)
11	0.75 (0.03)	−0.10 (0.09)	−0.63 (0.06)	−0.47 (0.18)	7.47 (0.04)	0.57 (0.50)	1832.21 (697.57)
12	0.76 (0.03)	−0.04 (0.09)	−0.56 (0.08)	−0.74 (0.16)	7.34 (0.03)	0.60 (049)	1765.27 (748.30)
13	0.72 (0.03)	0.05 (0.08)	−0.46 (0.07)	−1.12 (0.16)	7.40 (0.03)	0.67 (0.47)	1806.12 (705.39)
14	0.70 (0.04)	−0.02 (0.08)	−0.48 (0.08)	−0.91 (0.16)	7.38 (0.03)	0.64 (0.48)	1751.87 (667.73)
15	0.65 (0.04)	0.15 (0.06)	−0.54 (0.07)	−0.77 (0.13)	7.42 (0.03)	0.63 (0.48)	1885.57 (755.44)
16	0.72 (0.04)	0.12 (0.08)	−0.59 (0.06)	−0.70 (0.14)	7.37 (0.04)	0.61 (0.49)	1791.09 (712.16)
17	0.46 (0.05)	0.21 (0.07)	−0.40 (0.10)	−0.04 (0.10)	7.58 (0.04)	0.51 (0.50)	2233.96 (806.44)
Silly Sounds Stroop							
1	0.63 (0.04)	0.30 (0.09)	−0.29 (0.11)	−0.27 (0.14)	7.78 (0.02)	0.55 (0.50)	2386.62 (540.27)
2	0.47 (0.05)	−0.11 (0.06)	0.14 (0.09)	0.04 (0.12)	7.56 (0.04)	0.49 (0.50)	2141.41 (710.89)
3	0.59 (0.04)	−0.02 (0.08)	−0.27 (0.13)	−0.60 (0.12)	7.62 (0.03)	0.61 (0.49)	2157.56 (641.32)
4	0.68 (0.04)	−0.07 (0.08)	−0.45 (0.11)	−0.74 (0.14)	7.68 (0.02)	0.62 (0.49)	2259.78 (533.03)
5	0.72 (0.04)	0.10 (0.09)	−0.51 (0.12)	−0.29 (0.16)	7.62 (0.03)	0.54 (0.50)	2099.12 (582.47)
6	0.67 (0.05)	0.08 (0.09)	−0.57 (0.16)	−0.79 (0.16)	7.58 (0.03)	0.63 (0.48)	2064.98 (606.59)
7	0.49 (0.04)	−0.20 (0.08)	−0.47 (0.21)	−0.22 (0.12)	7.55 (0.04)	0.54 (0.50)	2124.78 (669.99)
8	0.51 (0.06)	−0.21 (0.05)	−0.24 (0.13)	−0.56 (0.11)	7.49 (0.03)	0.61 (0.49)	2005.69 (699.03)
9	0.55 (0.05)	−0.11 (0.06)	−0.39 (0.19)	−0.44 (0.11)	7.66 (0.02)	0.58 (0.49)	2211.77 (564.99)
10	0.67 (0.04)	−0.16 (0.08)	−0.34 (0.14)	−0.70 (0.16)	7.55 (0.04)	0.61 (0.49)	2095.67 (597.19)
11	0.58 (0.05)	−0.12 (0.07)	−0.29 (0.17)	−0.38 (0.14)	7.59 (0.03)	0.57 (0.50)	2134.07 (603.55)
12	0.54 (0.05)	−0.19 (0.07)	−0.36 (0.14)	−0.37 (0.13)	7.53 (0.04)	0.57 (0.50)	2060.94 (650.93)
13	0.61 (0.05)	0.00 (0.08)	−0.29 (0.13)	−0.62 (0.15)	7.62 (0.03)	0.61 (0.49)	2168.09 (609.88)
14	0.68 (0.04)	−0.15 (0.07)	−0.32 (0.11)	−0.74 (0.15)	7.52 (0.03)	0.62 (0.49)	2026.48 (605.57)
15	0.64 (0.05)	−0.04 (0.08)	−0.54 (0.17)	−0.45 (0.15)	7.56 (0.03)	0.58 (0.49)	2026.75 (593.73)
16	0.72 (0.04)	−0.25 (0.06)	−0.66 (0.13)	−0.55 (0.14)	7.48 (0.04)	0.58 (0.49)	2025.52 (626.41)
17	0.62 (0.05)	−0.10 (0.08)	−0.51 (0.16)	−0.44 (0.12)	7.63 (0.02)	0.58 (0.49)	2188.82 (561.27)
Animal Go/No-Go							
1	0.58 (0.06)	0.52 (0.13)	0.07 (0.24)	−2.23 (0.20)	6.82 (0.10)	0.85 (0.36)	1424.45 (766.49)
2	0.72 (0.05)	0.45 (0.15)	−0.56 (0.15)	−2.37 (0.24)	6.72 (0.18)	0.83 (0.38)	1471.85 (881.36)
3	0.71 (0.06)	0.13 (0.13)	−0.30 (0.18)	−2.40 (0.25)	7.05 (0.09)	0.83 (0.37)	1408.80 (659.57)
4	0.76 (0.05)	0.26 (0.16)	−0.32 (0.23)	−2.45 (0.35)	6.91 (0.12)	0.82 (0.39)	1388.88 (684.38)
5	0.61 (0.06)	0.14 (0.10)	−0.49 (0.27)	−1.52 (0.18)	7.08 (0.05)	0.76 (0.43)	1427.88 (701.40)
6	0.73 (0.05)	0.26 (0.15)	−0.36 (0.22)	−2.28 (0.25)	6.90 (0.13)	0.81 (0.39)	1371.10 (678.55)
7	0.62 (0.06)	0.35 (0.13)	−0.31 (0.21)	−1.97 (0.21)	6.82 (0.10)	0.81 (0.39)	1353.52 (796.40)
8	0.58 (0.07)	0.11 (0.11)	−0.68 (0.17)	−1.58 (0.15)	7.06 (0.08)	0.77 (0.42)	1401.98 (761.59)

Note. *SE* = standard error; RT = reaction time; IC = inhibitory control. All factor loadings are standardized. The leftmost column indexes the number of items on each task. The two rightmost columns show item-level descriptive statistics.

Silly Sounds Stroop

This Stroop-like IC task consisted of 17 items. Each item displayed pictures of a dog and cat (the left–right placement on the screen varied across trials) and presented the sound of either a dog barking or

cat meowing. Children were instructed to touch the picture of the animal that did not make the sound (e.g., touch the cat when hearing a dog bark). Each item was presented for 3,500 milliseconds irrespective of the speed of a child's response. The accuracy of each response was recorded, as was the associated RT for each item (RT responses less than 300 milliseconds were deemed too fast to be valid and set to missing). All 17 items were used to index task performance. On average, children completed 58% of items correctly (*N* = 591, *M* = 0.58, *SD* = 0.27, range = 0.00-1.00). For each item that a child answered correctly, the RT was recorded; RTs for incorrect responses were structurally missing. The two rightmost columns of Table 1 show the item-level percent correct and average RT for correct responses, respectively.

Animal Go/No-Go

This go/no-go IC task consisted of 40 items. Individual pictures of animals were presented, and children were instructed to touch a centrally located "button" on their screen every time that they saw an animal (the "go" response) except when that animal was a pig (the "no-go" response). Each item was presented for 4,000 milliseconds irrespective of the speed of a child's response. The accuracy and RT of each response was recorded. In contrast to the other two tasks, correct responses to the items that required IC involved withholding a response (i.e., not touching the button during the no-go items). As such, RT was only obtained for incorrect responses (RT responses less than 300 milliseconds were deemed too fast to be valid and set to missing). Only the no-go items were used to index task performance, because these are the items that require IC. On average, children completed 81% of the no-go items correctly (*N* = 563, *M* = 0.81, *SD* = 0.24, range = 0.00-1.00). For each item that a child answered incorrectly, the RT was recorded; RTs for correct responses were structurally missing. The two rightmost columns of Table 1 show the item-level percent correct and average RT for incorrect responses, respectively.

Analytic Plan

To examine the degree to which RT can inform performance on the IC tasks above and beyond accuracy, a series of unidimensional and bifactor-like item-level factor analysis models were fit using M*plus* 7.3 (Muthén & Muthén, 2012). To establish a baseline model, unidimensional binary item-level factor analysis models were fit using only item accuracies as indicators of IC abilities. Factor scores (response pattern based *expected a posterior* scores) and their *SE*s were retained for subsequent analysis. For each unidimensional model fitting, two estimators were used: mean and variance adjusted weighted least squares (WLSMV) and maximum likelihood (ML) with robust *SE*s to account for nonnormality in the log transformed RTs. Because WLSMV is a limited information estimator, it yields measures of absolute model fit (i.e., root mean square error of approximation [RMSEA], comparative fit index [CFI], etc.) that are unavailable with ML; however, ML tends to provide more stable parameter estimates in the presence of many items, and, unlike WLSMV, it uses all of the information available in the response patterns (Wirth & Edwards, 2007). As a full information estimator, ML is also well-equipped to accommodate missing observations, as data from partially complete cases are used in parameter estimation. The use of an estimator that can handle missing data is particularly important in this case, as by design, there was a nontrivial amount of missing RT data across the three tasks. For these reasons, WLSMV was used only to obtain information about overall model fit; ML was used to estimate all parameters. To account for the nesting of children within schools, school was used as a cluster variable in all analyses.

The IC ability and speed factors (see Figure 1) were constrained to be orthogonal, and for model identification, the mean and variance for both factors were fixed to 0 and 1, respectively. A binary

item-level factor model was used to model the influence of the IC ability factor on item accuracies; a linear normal factor model was used to describe the influence of the general IC and specific speed factors on the log transformed RTs. All factor loadings were tested for statistical significance using an alpha level of .05. To evaluate the incremental value of using item-level RT above and beyond item-level accuracy, factor scores and *SE*s for the unidimensional IC ability factor were compared with factor scores and *SE*s for the IC ability factor that also incorporates RT information. Graphical methods were used to determine the degree to which including RT information improves the precision (i.e., decreases the *SE*s) of the IC ability scores. Specifically, *SE*s were plotted as a function of factor scores for both the unidimensional IC ability factor and the IC ability factor from the bifactor-like model.

Results

Model Fit and Parameter Estimates

The unidimensional model using only accuracy information fit the data well for two of the three tasks, Silly Sounds Stroop: RMSEA = 0.06, CFI = 0.93, Tucker–Lewis index (TLI) = 0.92; Animal Go/No-Go: RMSEA = 0.00, CFI = 1.00, TLI = 1.00; however, unidimensional model fit for the Spatial Conflict Arrows task was mediocre (Spatial Conflict Arrows: RMSEA = 0.09, CFI = 0.91, TLI = 0.90). The slightly worse fit for the Spatial Conflict Arrows task is likely due to some mild local dependence induced by the "blocking" of many of the incongruent items—that is, an incongruent item that falls immediately after another incongruent item is more likely to be answered correctly than an incongruent item that falls immediately after a congruent item. Because the goal of this research was to examine the contribution of RT information above and beyond accuracy and not to perform a thorough psychometric evaluation of the tasks, the fit of the Spatial Conflict Arrows task was considered adequate to move forward with other analyses. After achieving acceptable model fit based on accuracy data alone, the proposed bifactor-like models incorporating RT information in addition to accuracy were fit to the data for each task.

The standardized factor loadings, thresholds, and intercepts, as well as *SE*s of estimates, can be found in Table 1. The accuracy thresholds provide a measure of difficulty that can be interpreted on the scale of the latent variable (i.e., standard scores). For example, an item with an accuracy threshold of 0 indicates that someone of average IC ability (factor score = 0) has a 50% probability of answering the item correctly. The smaller or more negative the threshold, the easier the item. The threshold estimates in Table 1 indicate that even within a single task, there is variability in item difficulty.

Of particular interest was the strength of the relations (i.e., standardized factor loadings) between item-level RT and IC ability relative to speed. This information is provided by the parameter estimates in Table 1, but it is more easily seen in graphical form. Figure 2 displays boxplots of the standardized factor loading estimates obtained from fitting the bifactor-like model to the data from each of the three tasks. Three sets of standardized estimates are shown: the loadings of item accuracy on IC ability, the loadings of RT on IC ability, and the loadings of RT on the specific speed factor. These boxplots reveal three main findings. First, IC ability is primarily informed by accuracy, not RT. This is particularly evident for Spatial Conflict Arrows, where the standardized factor loadings of accuracy on IC ability range from 0.32 to 0.76, while the loadings of RT on IC ability range from 0.18 to 0.30 (and tend not to be significantly different from 0 at a 0.05 level of significance). For the Silly Sounds Stroop and Animal Go/No-Go tasks, IC ability is at least partially informed by RT. While the standardized factor loadings of accuracy on IC (Silly Sounds Stroop: 0.47-0.72; Animal Go/No-Go: 0.58-0.76) are considerably larger than those of RT on IC ability (Silly Sounds Stroop: 0.25-0.30; Animal Go/No-Go: 0.11-0.52), several of

the loadings of RT on IC ability are statistically significant at the 0.05 level, albeit of small absolute magnitude.

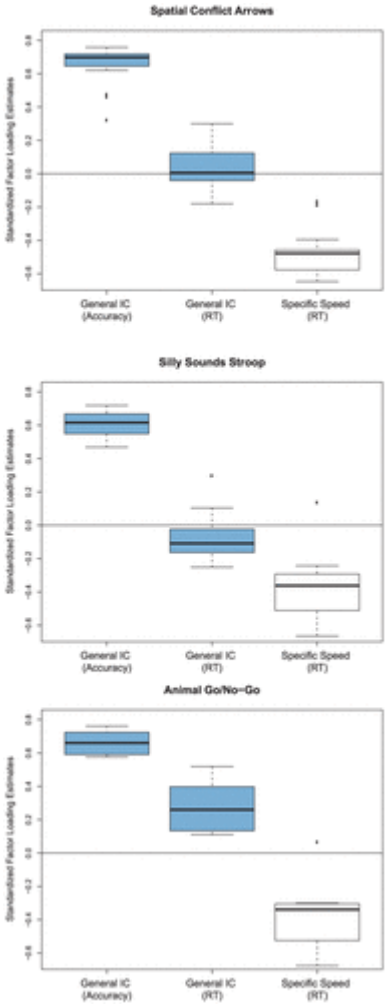

Figure 2. Boxplots of standardized factor loadings.

Note. RT = reaction time; IC = inhibitory control. Within the plot for each task, from left to right: (a) standardized factor loading of item-level accuracy on IC ability, (b) standardized factor loading of item-level RT on IC ability, and (c) standardized factor loading of item-level RT on speed.

Second, most of the information in RT is informing speed, not IC ability. This is observed across all three tasks, where the loadings of RT on the specific speed factor tend to be larger in magnitude than the loadings of RT on IC ability. Importantly, the speed factor in this model represents residualized individual differences that are specific only to RT, above and beyond any individual differences in IC ability. The large loadings of RT on speed relative to RT on IC ability suggests that there is substantial variability in RTs that is not explained by IC ability. The exception is for Animal Go/No-Go; for this task, some of the factor loadings of RT on IC are actually larger than those of RT on speed, suggesting that RT informs both speed and IC ability.

Third, the relative contribution of item-level RT to IC ability varies across the three tasks. For Spatial Conflict Arrows, the contribution of RT to IC ability is minimal. For Silly Sounds Stroop, the contribution is small but statistically significant. Children who are higher on IC tend to answer more items correctly (positive relationship between IC and accuracy), and they also tend to answer the items more quickly

(negative relationship between IC and RT). The relative contribution of RT to the measurement of IC is most noteworthy for Animal Go/No-Go. As expected, those who are higher on IC ability tend to answer more items correctly; however, for this task, it is slower RTs that are associated with higher levels of IC, shown in the positive factor loadings of RT on IC ability. This finding is not unexpected. The correct response for these items is the withholding of a response, and RT information is only available for children who answered the item incorrectly. Thus, children who answer the items incorrectly and quickly tend to exhibit the lowest levels of IC ability.

Factor Scores and Measurement Precision

An additional goal of this research was to evaluate the degree to which the use of item-level RT improves measurement precision above and beyond item-level accuracy. Figure 3 comprises three plots that highlight the contribution of RT to measurement precision for each task. Levels of IC ability (i.e., the observed factor scores) are shown along the *x*-axis; *SE*s of the factor score estimates are shown along the *y*-axis. The factor scores and *SE*s in this figure correspond to either (a) the unidimensional IC ability factor that is based solely on response accuracy or (b) the bifactor-like IC ability factor that is based on both response accuracy and RT information. The size of the circle or square is proportional to the number of children with a given score along the *x*-axis. Consistent with the standardized factor loading estimates, the contribution of RT to the precision of IC scores varies across the three tasks. For Spatial Conflict Arrows, the inclusion of RT does not yield smaller *SE*s for the IC ability factor scores. This can be seen in the close overlap between the grey circles, which reflect unidimensional IC ability scores based only on accuracy, and the black squares, which reflect IC ability scores from the bifactor-like model that incorporates RT information. For the two remaining tasks, the inclusion of RT information does, to some degree, reduce the *SE*s of the IC ability factor scores. While the effect is small for Silly Sounds Stroop, the *SE*s are reduced, particularly for children who are higher on IC ability. However, the reduction in *SE*s is most salient for Animal Go/No-Go; for below-average factor scores, the inclusion of RT information leads to a decrease in *SE*s. This is particularly true for scores that correspond to very low levels of IC ability, where *SE*s decrease by as much as 17%.

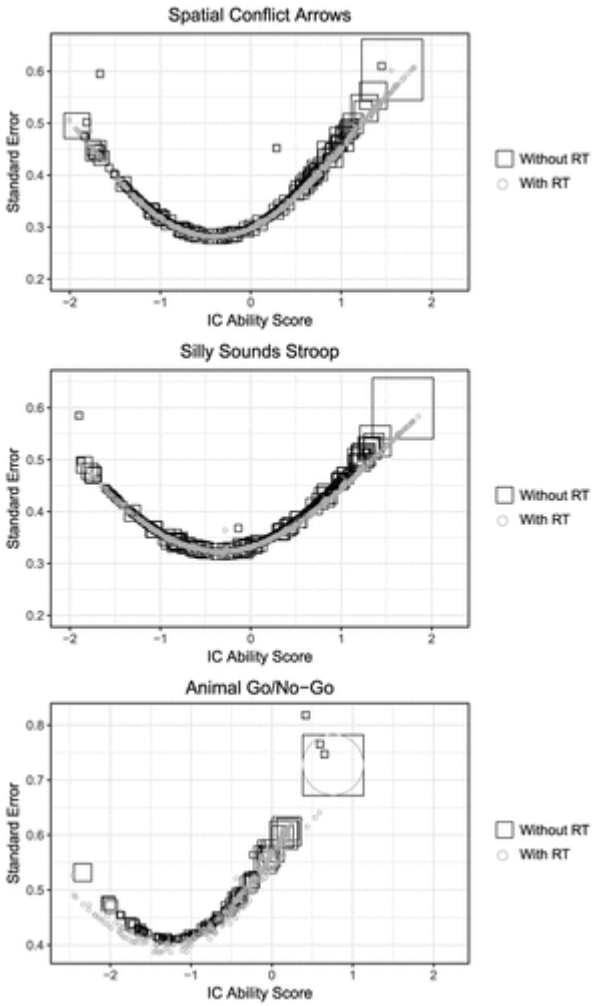

Figure 3. Standard errors (*y*-axis) as a function of IC ability factor scores (*x*-axis).

Note. RT = reaction time; IC = inhibitory control. Black squares: Unidimensional IC ability scores based on accuracy alone. Grey circles: Bifactor IC ability scores based on accuracy and RT. Across all levels of IC ability, the size of the shape is proportional to the number of children with a given score (i.e., large squares indicate floor or ceiling effects that manifest when only item-level accuracy is considered, small circles reflect the diffusion of scores that occurs with the incorporation of RT information).

In describing the relationship between IC ability scores and their associated *SE*s, it is important to note that each task has a small number of bivariate outliers that have larger than expected *SE*s for a given factor score. These outliers, which can be identified graphically in Figure 3, correspond to individuals with missing item-level accuracy information. The latent variable modeling framework allows for the response pattern-based scoring of all individuals, regardless of whether the response pattern is complete; however, the uncertainty of a score computed in the presence of missing data is reflected in larger *SE*s. For two individuals with the same factor score, the person with missing accuracy information has a larger *SE* than the person with a complete response pattern. This is why a small number of individuals have *SE*s that stand out relative to their associated factor score. For example, for Silly Sounds Stroop, the outlier in the upper left corner of the plot (IC ability = −1.89, *SE* = 0.59) is missing accuracy data for six items, whereas the less noticeable outlier in the middle of the plot (IC ability = −0.27, *SE* = 0.37) is only missing accuracy data for three items.

While the inclusion of RT information does not always yield smaller *SE*s, RT information does help in reducing the floor and ceiling effects that are present when only item-level accuracy is used as an indicator of IC ability, and this result is consistent across all three tasks. When only accuracy information is considered (i.e., the unidimensional model), 52 children earned the maximum score on Spatial Conflict Arrows, shown with the largest black square that is located at the factor score value of 1.56. When RT information supplements accuracy information, these 52 children spread out across a larger range of scores, shown with the several small grey dots ranging from 1.26 to 1.82. A similar effect is observed for Silly Sounds Stroop. The 37 children earning the maximum accuracy-based score of 1.68 are spread across a factor score range of 1.33 to 1.86 when RT information is used to supplement accuracy information. Unlike the other two tasks, for Animal Go/No-Go the spreading out of scores occurs only for the nine children earning the minimum accuracy-based score of −2.34: When RT information is included, the scores for these children spread out to a range of −2.51 to −2.12. The diffusion of scores is observed only at the lower tail of the distribution because RT information was not available for children answering the items correctly. Thus, RT reduces the presence of the floor effect but not the ceiling effect for this task.

In summary, the inclusion of RT information provides two main measurement benefits that are task dependent. For Spatial Conflict Arrows and Silly Sounds Stroop, the RTs primarily achieve more fine-grained measurement of IC ability, particularly at the extreme levels where floor and ceiling effects tend to exist when only accuracy information is considered. These two tasks also benefit from a slight reduction in *SE*s at high levels of IC ability, although this effect is minimal. For Animal Go/No-Go, the main benefit of including RT is the improvement in measurement precision only at low ends of IC ability.

Discussion

The overall objective of this study was to test to what extent item-level RT data could be used to supplement item-level accuracy data for purposes of scoring IC tasks. A related objective was to determine the incremental value of incorporating RT information into IC factor scores. Results demonstrated that the inclusion of RTs helped both improve measurement precision and reduce floor and ceiling effects relative to task scores that were based solely on the accuracy of responses. These findings have implications for studies that use IC factor scores as predictors, mediators, moderators, or outcome measures of EF—more information is contained in scores that include RT than scores that do not.

We demonstrated a relatively simple analytic approach for jointly modeling item-level accuracy and RT information for purposes of scoring IC tasks. The bifactor-like model used here aligns closely with other psychometric models that were developed to jointly model accuracy and RT; however, unlike other models, it facilitates the derivation of a single task score that combines item accuracy and only that portion of RT variation that is indicative of IC abilities. Notably, this model can be estimated using standard latent variable modeling software, which makes it accessible to a broad audience of researchers, and it leverages the strengths of general latent variable modeling methods (e.g., accommodates missing data; correction for measurement error). This modeling approach also provides a framework for testing many of the assumptions of methods that are currently used for combining accuracy and RT data (e.g., the DCCS and Flanker tasks as implemented in the NIH Toolbox), and in some cases, shows that the tenability of those assumptions is questionable. For example, comparing the relative size of the standardized factor loadings on the IC ability shows that accuracy contributed

more to IC ability than did RTs in our preschool-aged sample. This has been observed in other preschool-aged samples (e.g., Davidson, Amso, Anderson, & Diamond, 2006) and contradicts the idea that accuracy and RT are equally important indicators of IC. Furthermore, the results also suggest that even though these tasks are designed to have items of equal difficulty, items can still vary in their difficulty levels within the same task.

Our results demonstrated the merits of jointly modeling accuracy and RT data especially compared with tasks that have typically been scored exclusively using accuracy information. For Silly Sounds Stroop and Animal Go/No-Go, there was evidence that item-level RT information improved the measurement of IC abilities. The reduction in *SE*s was most noteworthy for Animal Go/No-Go, where at low levels of IC ability, RT information decreased *SE*s by as much as 17%. We surmise that RTs were particularly informative for this task for two reasons. First, this was the easiest task of the three, with an average accuracy rate of 80%: Most respondents were able to answer these items correctly; hence, more RT information was structurally missing in this task. Consequently, RTs are only able to help with measurement precision for individuals who tend to answer the items incorrectly—those at low levels of IC ability. Second, and related to the first point, RTs may be especially useful for tasks where the correct response is the inhibition of a response. Arguably, someone who answers the item incorrectly but takes more time before responding has stronger inhibitory abilities than someone who immediately answers the item incorrectly on impulse. Both the ease and response format (i.e., the withholding of a response is the correct answer) of go/no-go tasks may make RTs especially useful in improving the measurement precision of scores that fall at the lower ends of the ability continuum.

For all three tasks, the addition of RT data helped minimize floor or ceiling effects, depending on the response format of the task. For Spatial Conflict Arrows and Silly Sounds Stroop, RT information allows for finer grained measurement primarily at high levels of IC ability where there were previously ceiling effects; this result is consistent with the expectation that individuals who answer the items correctly, and therefore also have RTs associated with their responses, are of higher ability levels. For Animal Go/No-Go, the opposite effect is observed; because RTs are only recorded for incorrect responses, finer grained measurement tends to occur only at lower levels of IC ability. The inclusion of RT information does not diffuse the ceiling effect for this task.

It is important to note that these benefits would not have been evident had we asked a more simplistic question of how highly correlated accuracy-only versus accuracy+RT scores were (all scores were highly correlated, *r*s = .99). Although incorporating RT information did not change the rank order of children's performance, it did improve measurement precision, particularly for the 6% to 8% of children who scored at the extremes and would have otherwise all received the same score. Moreover, although improvements in the precision of measurement were of modest magnitude and varied by task, the RT data were "free," in that they were automatically recorded and did not require any additional burden on participants. Improving task scoring without increasing test burden is particularly important in assessments of young children. All of these merits suggest that RT information can offer some contribution to the assessment of IC ability; however, we emphasize that in the joint analysis of item-level accuracy and RT data, it is accuracy that remains the primary indicator of IC abilities.

While the focus of this study was on whether RT data could be leveraged to improve the measurement of IC abilities in young children, the results of this study have further reaching implications. The ability to jointly model accuracy and RT data for purposes of task scoring opens the prospect of using a common set of tasks over longer developmental periods than has typically been possible. The results of this study emphasize that much of the variation in item-level RT data is indicative of individual

differences in speed of processing—not IC. Many traditional approaches to scoring IC tasks that use RT data conflate IC abilities with a general processing speed. This has the potential to undermine program evaluation efforts (e.g., evidence for or against intervention effects are ambiguous because it is unclear whether they represent IC or processing speed) and to characterize normative developmental changes in IC (e.g., some of the apparent age-related improvements in IC represent improvements in processing speed).

Given that these IC tasks are already routinely administered by computer or tablet, it may be of potential benefit to consider whether they could be modified for use within a computer adaptive test (CAT) framework; the availability of both accuracy and RT data makes this approach especially appealing. Within the CAT framework, items are administered from an item bank such that, conditional on the respondent's current estimated location on the latent variable (e.g., IC ability level), each successively administered item is maximally informative (Wainer & Mislevy, 1990)—that is, the item is selected to yield the greatest reduction in the *SE* of the current ability estimate. Based on the response to each successively administered item, the estimated ability level is then updated. If the respondent answers that item incorrectly, an easier item is administered, whereas if the respondent answers that item correctly, a more difficult item is administered. Once the *SE* of the current ability estimate drops below a predetermined threshold (i.e., so-called "target precision"), the test stops and a score is assigned to the respondent. The idea of developing an item bank for IC tasks is unusual because they typically involved the repeated administration of the same set of items (e.g., dog and cat sound associations in our Silly Sounds Stroop task). However, the results of this study demonstrate that the placement of items within the task have an impact on their difficulty and discrimination levels. The results of this study also suggest that item-level RT is another indicator that could be used to inform person ability level. Although not considered here, the relative timing of item presentation is yet another potential source of variation that could affect the accuracy and RT of item responses. An important direction for future research involves determining whether and how CAT methods might be used to result in shorter IC tasks without sacrificing measurement precision.

A key limitation of our study is that we present only internal evidence that IC is better measured using both accuracy and RT information. Future studies should provide tests of the predictive validity of accuracy-only versus accuracy+RT based scores. For example, it would be useful to test whether RT-enhanced scores better predict school outcomes. Results of such studies would provide a stronger test of the merits of using the bifactor-like modeling approach that was used here. Future research would also benefit from the inclusion of children across a wide age range. Our results suggest that item-level RT data make an important but limited contribution to the measurement of IC abilities among young children. It remains an open question whether and to what extent RT enhance the measurement of IC abilities in older children. As children's accuracy of performance asymptotes with advancing age, item-level RT data may become comparatively more informative of IC ability.

Authors' Note

The views expressed in this article are those of the authors and do not necessarily represent the opinions or position of the Institute of Educational Sciences.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was supported by Institute of Educational Sciences Grant R324A120033.

References

Blair, C., Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. *Development and Psychopathology*, 20, 899-911.

Blair, C., Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, 66, 711-731. doi:10.1146/annurev-psych-010814-015221

Cai, L., Yang, J. S., Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16, 221-248. doi:10.1037/a0023350

Camerota, M., Willoughby, M. T., Kuhn, L. J., Blair, C. B. (2016). The Child Executive Functioning Inventory (CHEXI): Factor structure, measurement invariance, and correlates in US preschoolers. *Child Neuropsychology*. Advance online publication. doi:10.1080/09297049.2016.1247795

Carlson, S. A. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*, 28, 595-616.

Carlson, S. M., Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, 72, 1032-1053.

Davidson, M. C., Amso, D., Anderson, L. C., Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, 44, 2037-2078. doi:10.1016/j.neuropsychologia.2006.02.006

De Luca, C. R., Leventer, R. J. (2008). Developmental trajectories of executive functions across the life span. In Anderson, V., Jacobs, R., Anderson, P. J. (Eds.), *Executive functions and the frontal lobes* (pp. 22-56). New York, NY: Taylor & Francis.

De Luca, C. R., Wood, S. J., Anderson, V., Buchanan, J. A., Proffitt, T. M., Mahony, K., Pantelis, C. (2003). Normative data from the CANTAB. I: Development of executive function over the lifespan. *Journal of Clinical and Experimental Neuropsychology*, 25, 242-254. doi:10.1076/jcen.25.2.242.13639

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64, 135-168. doi:10.1146/annurev-psych-113011-143750

Garon, N., Smith, I. M., Bryson, S. E. (2014). A novel executive function battery for preschoolers: Sensitivity to age differences. *Child Neuropsychology*, 20, 713-736. doi:10.1080/09297049.2013.857650

Knudsen, E. I., Heckman, J. J., Cameron, J. L., Shonkoff, J. P. (2006). Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 10155-10162. doi:10.1073/pnas.0600888103

Molenaar, D., Tuerlinckx, F., van der Maas, H. L. J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, 50, 56-74. doi:10.1080/00273171.2014.962684

Muthén, L. K., Muthén, B. O. (1998-2012). *Mplus User's Guide: Statistical Analysis with Latent Variables* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin*, 126, 220-246.

Petersen, I. T., Hoyniak, C. P., McQuillan, M. E., Bates, J. E., Staples, A. D. (2016). Measuring the development of inhibitory control: The challenge of heterotypic continuity. *Developmental Review*, 40, 25-71. doi:10.1016/j.dr.2016.02.001

Shanmugan, S., Satterthwaite, T. D. (2016). Neural markers of the development of executive function: Relevance for education. *Current Opinion in Behavioral Sciences*, 10, 7-13. doi:10.1016/j.cobeha.2016.04.007

Ursache, A., Blair, C., Raver, C. C. (2012). The promotion of self-regulation as a means of enhancing school readiness and early achievement in children at risk for school failure. *Child Development Perspectives*, 6, 122-128. doi:10.1111/j.1750-8606.2011.00209.x

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Edcuational Measurement*, 46, 247-272.

van der Linden, W. J. (2016). Lognormal response-time model. In van der Linden, W. J. (Ed.), *Handbook of item response theory* (Vol. 1, pp. 261-282). New York, NY: Taylor & Francis.

Wainer, H., Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In Wainer, H., Dorans, N., Flaugher, R., Green, B., Mislevy, R., Steinberg, L., Thissen, D. (Eds.), *Computerized adaptive testing: A primer* (pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum.

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

Willoughby, M., Kupersmidt, J., Voegler-Lee, M., Bryant, D. (2011). Contributions of hot and cool self-regulation to preschool disruptive behavior and academic achievement. *Developmental Neuropsychology*, 36, 162-180. doi:10.1080/87565641.2010.549980

Willoughby, M. T., Blair, C. B. (2016). Longitudinal measurement of executive function in preschoolers. In Griffin, J., Freund, L., McCardle, P. (Eds.), *Executive function in preschool age children: Integrating measurement, neurodevelopment and translational research* (pp. 91-113). Washington, DC: American Psychological Association.

Willoughby, M. T., Blair, C. B., & Family Life Project Investigators. (2016). Measuring executive function in early childhood: A case for formative measurement. *Psychological Assessment*, 28, 319-330. doi:10.1037/pas0000152

Willoughby, M. T., Blair, C. B., Wirth, R. J., Greenberg, M., & Family Life Project Investigators. (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.

Willoughby, M. T., Blair, C. B., Wirth, R. J., Greenberg, M., & Family Life Project Investigators. (2012). The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment*, 24, 226-239.

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.

Wirth, R. J., Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12(1), 58-79.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beamont, J. L., Weintraub, S. (2013). II. NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development*, 78(4), 16-33.