AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

© 2019 American Psychological Association
1040-3590/19/$12.00

2019, Vol. 31, No. 8, 995–1005
http://dx.doi.org/10.1037/pas0000725

Speed and Accuracy on the Hearts and Flowers Task Interact to Predict Child Outcomes

Marie Camerota and Michael T. Willoughby
RTI International, Research Triangle Park, North Carolina

Clancy B. Blair
New York University

The current study tests whether accuracy and reaction time (RT) on the Hearts and Flowers (HF) task, a common assessment tool used across wide age ranges, can be leveraged as joint indicators of child executive function (EF) ability. Although previous studies have tended to use accuracy or RT, either alone or as separate indicators, one open question is whether these 2 metrics can be yoked together to enhance our measurement of EF ability. We test this question using HF data collected from first-grade children who participated in the Family Life Project. Specifically, we model the independent and interactive effects of HF accuracy and RT on several criterion outcomes representing child academic and behavioral competence. Our findings indicate that among early-elementary-aged children, accuracy and RT interact in the prediction of child outcomes, with RT being a more informative index of EF ability for children who perform at high levels of accuracy. The main effect of accuracy remained significant in the presence of these interactive effects. This pattern of findings was similar for different task blocks (i.e., mixed, flower-only) and for different child outcome domains (i.e., academic, behavioral). Our finding of an interaction between accuracy and RT contributes to a growing literature that attempts to jointly consider accuracy and RT as indicators of underlying ability, which has important implications for how EF task scores are constructed and interpreted.

Public Significance Statement
This study finds that accuracy and reaction time (RT) on executive function (EF) tasks may both provide useful information about a child's EF ability. In addition, RT may be a more informative indicator of EF ability for children who perform at high levels of accuracy.

Keywords: executive function, accuracy, RT, task scoring

In recent years, there has been increased scientific interest in the construct of executive function (EF). EF is a multifaceted construct that encapsulates the set of higher order cognitive abilities that support planful, goal-directed behavior (Blair & Ursache, 2011). In children, EF is thought to be comprised of three related, yet separable, abilities: working memory, inhibitory control, and cognitive flexibility (Miyake et al., 2000). Researchers across psychological disciplines are interested in measuring EF because it predicts a wide variety of outcomes, including school readiness (e.g., Blair & Razza, 2007), risk of psychopathology (for a review, see Snyder, Miyake, & Hankin, 2015), occupational functioning (e.g., Miller, Nevado-Montenegro, & Hinshaw, 2012), and quality of life (e.g., Davis, Marra, Najafzadeh, & Liu-Ambrose, 2010). Put another way, EF seems necessary for humans to flourish in all stages of life. Despite this, longitudinal studies examining the development of EF across the life span are relatively rare.

This article was published Online First April 29, 2019.

Marie Camerota and Michael T. Willoughby, Education and Workforce Development, RTI International, Research Triangle Park, North Carolina; Clancy B. Blair, Department of Applied Psychology, New York University.

Data collection for this study was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development P01 HD039667. Data analysis and writing for this study was supported by the Office of The Director, National Institutes of Health of the National Institutes of Health under Award UG3OD023332. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This study is part of the Family Life Project (https://flp.fpg.unc.edu/).

The Family Life Project Phase II Key investigators include Lynne Vernon-Feagans, The University of North Carolina at Chapel Hill; Mark T. Greenberg, The Pennsylvania State University; Clancy B. Blair, New York University; Margaret R. Burchinal, The University of North Carolina at Chapel Hill; Martha Cox, The University of North Carolina at Chapel Hill; Patricia T. Garrett-Peters, The University of North Carolina at Chapel Hill; Jennifer L. Frank, The Pennsylvania State University; W. Roger Mills-Koonce, University of North Carolina-Greensboro; and Michael T. Willoughby, RTI International.

Correspondence concerning this article should be addressed to Marie Camerota, Education and Workforce Development, RTI International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709. E-mail: mcamerota@rti.org

One reason for this is the lack of performance-based measures that can be used across different age groups. Even with adaptations, tasks developed for use with children are too easy for adults, and tasks developed for adults are too difficult for children. A second limitation is that, even among the relatively small number of tasks that can be used across wide age ranges, there is disagreement regarding optimal measures of performance for different age groups. For example, although there is a precedent for relying on accuracy of performance for younger children (Diamond, Barnett, Thomas, & Munro, 2007), reaction time (RT) is more often used for older children and adults (Diamond & Kirkham, 2005). However, there is little empirical guidance regarding when to make the switch from accuracy to RT as an index of performance. To complicate matters even further, stimulus presentation rate is adjusted downward to make a given task more difficult for older children and adults (Davidson, Amso, Anderson, & Diamond, 2006). This change means that it can again be difficult, if not impossible, to model performance on a given task across time.

The Hearts and Flowers (HF) task is one example of a task that has been used in early childhood, middle childhood, and adolescence. An adaptation of the Dots task originally developed by Davidson and colleagues (2006), the HF task requires participants to respond to one rule (press a key on the same side of the screen) when presented with a heart stimulus and to respond to a different rule (press a button on the opposite side of the screen) when presented with a flower stimulus. The task consists of three blocks containing either congruent (hearts only), incongruent (flowers only), or mixed (both hearts and flowers) trials. Because there are no executive demands inherent in the hearts block, it is often used as a control block from which to compare performance on the flower and mixed blocks. Commonly, performance on the flower block is used as an index of inhibitory control (Wright & Diamond, 2014) because it requires inhibiting a prepotent response (i.e., pressing the spatially congruent key). Performance on the mixed block is used as an index of cognitive flexibility, as it requires flexibly selecting between the two rules, depending on the stimulus. Although the task more intentionally taxes inhibitory control and cognitive flexibility, others have argued that working memory is inherently involved in all trials because of the need to hold two rules in mind throughout the different blocks of the task (Diamond, 2013).

The HF task has been shown to be appropriate for Ages 4 years through adulthood, although, as noted previously, a downward adjustment to the stimulus presentation time is needed to keep the task challenging for older children and adults (Davidson et al., 2006). Specifically, the recommended length of time for stimulus presentation is 2,500 ms for children 6 years and younger, whereas 750 ms is recommended for older children and adults. The HF task has been shown to demonstrate age-related shifts in accuracy and RT (Davidson et al., 2006), discriminative validity to distinguish between typically and atypically developing children (Edgin et al., 2010), and sensitivity to various types of educational and physical activity intervention programs (Diamond et al., 2007; Lakes et al., 2013; Schonert-Reichl et al., 2015).

Despite its many favorable properties, one difficulty of using this task, and others like it, is that there are multiple ways performance can be summarized. Some studies focus on performance within each of the three task blocks (e.g., hearts, flowers, and mixed). Others examine performance on all congruent (i.e., hearts) and incongruent (i.e., flowers) trials, regardless of task block. In addition to these different ways to group trials, there are at least three different indicators of performance: accuracy (i.e., percent correct), RT (i.e., mean RT in ms), and RT difference scores (e.g., mean RT on mixed trials – mean RT on hearts trials, mean RT on incongruent trials – mean RT on congruent trials). Thus, with five ways to group trials and three performance indicators, there are many choices for summarizing performance on the HF task. In the absence of clear guidance regarding the choice of developmentally appropriate measures, studies seem to arbitrarily choose some subset of accuracy, RT, and/or RT difference scores.

Some studies solely report findings pertaining to accuracy. For example, Diamond and colleagues (2007) found that preschool children who were randomly assigned to an EF-training curriculum (Tools of the Mind) exhibited greater gains in accuracy on the incongruent (e.g., flowers) block of the HF task compared with peers who participated in a literacy-based curriculum. Citing concerns over the young age of the children, RT differences were not assessed. On the other hand, some studies with older children report findings solely pertaining to RT. One study found that higher levels of state anxiety in elementary-aged children was related to poorer EF ability, as indexed by larger RT difference scores on the mixed trials (Ursache & Raver, 2014). The authors report that RT difference scores were used for all analyses because mean accuracy in the sample was high ($M = 85\%$) and lacked adequate variability ($SD = .16$). Finally, some studies report findings for accuracy and RT together. For example, Davidson and colleagues (2006) observed that among 4- to 13-year-olds, children were both slower and less accurate on incongruent (e.g., flowers) trials than on congruent (e.g., hearts) trials, likely because of the increased inhibitory demand of incongruent trials. They also observed that the differences in RT and accuracy among these two types of trials diminished with age, as inhibitory control ability increased.

Clearly, there are differences in the metrics that different studies use and report, even when using the same task. These differences make it difficult to compare findings across studies. Practically, the use of different metrics among different age groups of children (i.e., accuracy with younger children, RT or RT difference with older children) also makes it difficult to model change in EF ability across time. Among the few studies that have examined both accuracy and RT, these metrics have tended to be analyzed independently as predictors or outcomes (e.g., Davidson et al., 2006; Lakes et al., 2013; Wright & Diamond, 2014). One question that remains is whether accuracy and RT can be yoked together in a way that would yield additional information about child EF ability. This question is particularly interesting in the case of young children, for whom accuracy seems to be the preferred metric. The question of whether accuracy and RT on the HF task can be jointly used as indices of EF ability in young children constitutes the primary motivation for the current investigation.

Research using other task batteries has attempted to integrate accuracy and RT into single scores. For example, the NIH Toolbox consists of two EF tasks, the Dimensional Change Cart Sort (DCCS), and Flanker tasks. Motivated by their desire to create task scores that capture meaningful variability in performance for individuals from Age 3 through adulthood, a two-vector scoring approach was proposed (Zelazo et al., 2013). Specifically, individuals receive separate scores for accuracy and RT, which are transformed from their raw form such that they are both measured on a 0 to 5 scale. For participants who are highly accurate (i.e., 80% accuracy or higher), the RT score is added to the accuracy score, yielding a possible score that ranges from 0 to 10. For individuals who do not score at 80% accuracy or higher, their score is solely determined by their accuracy, and ranges from 0 to 5. More sophisticated, model-based approaches have also been proposed for integrating accuracy and RT (Magnus, Willoughby, Blair, & Kuhn, 2017; Molenaar, Tuerlinckx, & van der Maas, 2015). These approaches use factor analytic and item response theory approaches to estimate the joint contributions of accuracy and RT to cognitive ability. Although these models are general enough to be of use to a wide audience, practically, they are beyond the analytic grasp of most substantive researchers.

The current study tests a simpler approach, which can be implemented in a traditional linear regression framework. Rather than making a priori task scoring decisions that delineate the conditions under which RT becomes an informative measure of EF ability (i.e., at a specific child age or past a set level of accuracy), we test a series of models that empirically address this question. Using data from a sample of first-grade children, we model the independent and interactive effects of HF accuracy and RT on several criterion outcomes. Main effects of accuracy and RT are tested to determine whether accuracy and RT independently account for variance in outcomes, which would indicate that both metrics provide unique information about child EF ability. We also test interactive effects of accuracy and RT, to determine whether the utility of RT as an index of EF ability depends upon the child's overall level of accuracy. In this way, we are able to empirically demonstrate whether there is a certain threshold of accuracy that must be reached before RT becomes informative.

In order to test these models, we first needed to identify criterion measures that we expected to be robustly related to EF ability in childhood. Studies have consistently indicated that among preschool and early-elementary-aged children, better EF is implicated in school readiness (e.g., Blair, 2002), prosocial behavior (Bierman, Torres, Domitrovich, Welsh, & Gest, 2009; Smith-Donald, Raver, Hayes, & Richardson, 2007), and academic achievement (Brock, Rimm-Kaufman, Nathanson, & Grimm, 2009; Bull & Scerif, 2001; Willoughby, Kupersmidt, Voegler-Lee, & Bryant, 2011), with evidence that these links are stronger for math than for reading ability (Blair & Razza, 2007; Espy et al., 2004). On the other hand, EF deficits are commonly reported in children with different classes of behavior problems, including attention-deficit hyperactivity disorder, and other externalizing disorders (Pauli-Pott & Becker, 2011; Séguin, 2004; Willcutt, Doyle, Nigg, Faraone, & Pennington, 2005). Therefore, we selected three scales of academic achievement and three scales of behavioral functioning as outcome measures for the current investigation, given their expected relationship to EF task performance. Although these outcomes are not of substantive interest to the current investigation, they serve as important benchmarks for evaluating the independent and interactive efficacy of accuracy and RT as indicators of child EF ability.

In the absence of previous studies that have made joint use of HF accuracy and RT scores in elementary-aged children, we refrain from making strong directional hypotheses about how these metrics will independently and interactively relate to child outcomes. However, given the implicit assumption in the EF literature that RT is more informative at higher levels of accuracy (e.g., Zelazo et al., 2013), we predict that accuracy and RT may interact in the prediction of child outcomes, such that RT may be a stronger predictor of child outcomes for individuals who perform at higher levels of accuracy. We do not have explicit hypotheses about differences in these models based on task block (i.e., flowers, mixed) or outcome domain (i.e., academic, behavioral outcomes).

Method

Participants

The Family Life Project (FLP) is a longitudinal study of children and families residing in two regions of high rural poverty in North Carolina (NC) and Pennsylvania (PA). Families living in target counties were recruited using a stratified random sampling approach that yielded a representative sample of 1,292 families recruited over a 1-year period (September 2003 through September 2004). Low-income families in both states and African American families in NC were oversampled to ensure adequate power to test central research questions. Additional details about FLP sampling and recruitment procedures can be found in Vernon-Feagans and Cox (2013).

The current analyses include a subsample of children drawn from the larger FLP study. To be included in these analyses, we first considered any child who had HF data and first-grade outcome data ($N = 1{,}020$). We excluded an additional 60 children whose HF data were deemed invalid (criteria described below, in the Measures section). Thus, the final analysis sample contained 960 children ($M_{\text{age}} = 87.3$ months, $SD_{\text{age}} = 3.8$ months). In this subsample, 50% of children were male, 43% were African American, and 78% of families were considered poor (<200% of the poverty level) at recruitment. This subsample does not differ from the full FLP sample in terms of child race, gender, research site (NC or PA), or poverty status at recruitment. The relevant institutional review boards approved all data collection activities, including informed consent.

All analyses accounted for the complex sampling design and clustered school-based data by incorporating appropriate stratification weight and cluster variables.

Measures

Data for these analyses are drawn from one school visit and one home visit conducted when children were in their second year of formal schooling. For the majority (93%) of children in this study, their second year of school corresponded to the first grade. However, a small number of children were in kindergarten (7%) or second grade (<1%) because of grade retention or acceleration, respectively. For ease of interpretation, we refer to this data collection time point as the "first-grade visit" throughout the article. In the spring of the school year, children completed a number of tasks, including assessments of math and reading ability. The target child's lead/primary teacher completed questionnaires regarding the child's behavior, including ratings of children's social skills and behavior problems. At the home visit, parents and children completed individual and dyadic activities, including an assessment of child EF ability.

Executive function. EF was measured using the HF task (Davidson et al., 2006). On each trial, children were presented with a picture of a heart or a flower on one side of a laptop screen. They were instructed to press the keyboard button on the same side as the picture when the picture was a heart (congruent) but to press the keyboard button on the opposite side as the picture when the picture was a flower (incongruent). Children completed instructional and practice trials, which were repeatable up to three times, until children understood task demands. These practice trials were followed by 12 hearts-only trials, 12 flower-only trials, and 33 mixed trials. Stimuli were presented for up to 2,500 ms (depending on whether a response was made) and advanced to the next trial following an interstimulus interval of 1,000 ms. Accuracy and RT were measured for each individual trial. Anticipatory responses ($\mathrm{RT} < 200$ ms) were set to missing for both accuracy and RT metrics, as these responses occurred too fast to be in response to the stimulus. Consistent with previous studies (e.g., Ursache & Raver, 2014), only children who responded to at least 75% of trials were considered to have valid HF data. Additionally, the current analyses excluded HF data from children who performed below chance levels (i.e., overall task accuracy <50%). As mentioned above, applying these criteria led us to exclude data from 60 (out of 1,020; 6%) children.

Accuracy scores were calculated for each block and represented the proportion of correct responses (e.g., correct responses divided by sum of correct and incorrect responses). Similarly, the mean RT of correct responses was calculated for each block. Subsequently, we calculated two RT difference scores (ΔRT) by subtracting individuals' mean RT on the heart-only block from their mean RT on the flower-only block (i.e., $\mathrm{RT}_{\mathrm{flower}} - \mathrm{RT}_{\mathrm{heart}}$) and mixed block (i.e., $\mathrm{RT}_{\mathrm{mixed}} - \mathrm{RT}_{\mathrm{heart}}$). These ΔRT scores represented the slowing related to increased inhibitory control and shifting demands, respectively. Four measures (accuracy and ΔRT for flower-only and mixed blocks) were retained as focal predictor variables.

Academic outcomes. Academic outcomes included child math and reading ability, assessed using the Woodcock-Johnson III (WJ III) Tests of Achievement (Woodcock, McGrew, & Mather, 2001). The WJ III consists of a normed set of tests measuring cognitive abilities, scholastic aptitude, and academic achievement. Math ability was indexed using the Applied Problems subtest, in which children are asked to solve mathematical word problems. Reading ability was assessed using the Letter–Word Identification and Picture Vocabulary subtests of the WJ III. In the Letter-Word subtest, children are asked to identify letters and read words of increasing difficulty. In the Picture Vocabulary subtest, children are asked to name pictures. We used standard scores for each subtest as outcome measures. All tests within the WJ III Tests of Achievement have been shown to demonstrate high levels of reliability and validity (Woodcock et al., 2001).

Behavioral outcomes. Behavioral outcomes included teachers ratings of children's social skills and behavior problems using the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), a 25-item screener appropriate for children Ages 3 to 16 years. Teachers rated a series of items on a 3-point Likert scale (0 = *not true*, 1 = *somewhat true*, 2 = *certainly true*). We use three subscales in these analyses. The Conduct Problems subscale contains five items ($a = .80$) that measure child problem behaviors (sample item: "Often lies or cheats"). The Hyperactivity subscale contains five items ($a = .90$) that measure child inattention and hyperactivity (sample item: "Easily distracted, concentration wanders"). The Prosocial subscale contains five items ($a = .86$) that measure children's prosocial behavior and social skills (sample item: "Considerate of other people's feelings"). Mean scores were calculated for each subscale (range = 0 to 2).

Analytic Plan

Substantive questions were tested using a series of regression models predicting our six outcome measures (i.e., three WJ III and three SDQ scores). Our first test was whether mixed block accuracy and ΔRT predicted our criterion outcomes. This test investigated the contribution of cognitive flexibility trials, specifically, to child outcomes. The second test was whether flower-only block accuracy and ΔRT predicted these same outcomes. This test investigated the contribution of inhibitory control to our criterion measures. Finally, we tested whether mixed and flower-only accuracy and ΔRT jointly predicted child outcomes. Although the current investigation is not specifically concerned with the independent influence of cognitive flexibility and inhibitory control on child outcomes, we include these models because we acknowledge that substantive researchers may be interested in using these variables in this way.

For each condition (i.e., mixed block, flower-only block, both mixed and flower-only block), we estimated six models that included the direct effects of accuracy and ΔRT, as well as their interaction, in the prediction of child academic and behavioral outcomes. Raw accuracy and ΔRT variables were centered using weighted means prior to creating interaction terms. Weighted means were used for centering in order to account for the complex sampling design.

For models in which the interaction between accuracy and ΔRT is significant, we probe these interactions using regions of significance and simple slopes analyses. Simple slopes were calculated for low (25th percentile) and high (75th percentile) values of accuracy, to demonstrate the expected relationship between ΔRT and child outcomes at varying levels of task performance. Results from regions of significance analyses tell us at what level of accuracy ΔRT becomes a significant predictor of child outcomes. In cases in which the interaction between accuracy and ΔRT is not significant, we trim interaction terms prior to interpreting model coefficients. Trimming interaction terms in this way did not change the substantive model results. However, the results from models including the full set of predictor variables are available by request.

All analyses were conducted using PROC SURVEYREG in SAS 9.4. Robust standard errors for model coefficients were obtained using the CLUSTER command, to account for nesting of kids in classrooms (Huber, 1967). Missing data were handled using full-information maximum likelihood. The majority of cases ($n = 828$; 86%) had full data for all variables. Of those with missing data ($n = 132$; 14%), the mean number of missing variables was 3.1 ($SD = 0.68$) out of the 10 analysis variables. The majority of missingness was accounted for by 123 children who were missing SDQ ratings. Children with missing data did not differ from children without missing data on the basis of child race, gender, research site (NC or PA), or poverty status at recruitment.

Results

Descriptive Statistics

Unweighted means, standard deviations, and correlations for all study variables are presented in Table 1. Children demonstrated higher mean accuracy on flower-only trials (M = .89, SD = .17) compared with mixed trials (M = .81, SD = .16). Children also slowed their responding to a greater degree on mixed trials (M = 547.42, SD = 224.23) compared with flower-only trials (M = 244.68, SD = 206.09), indicating that mixed trials were more challenging than flower-only trials. Accuracy and ΔRT were positively correlated for both flower-only (r = .16, $p < .001$) and mixed (r = .21, $p < .001$) trials, indicating that children who were more accurate also tended to slow down more. There was also a negative correlation between children's RT on heart-only trials and their ΔRT on both flower-only ($r = -.28$, $p < .001$) and mixed ($r = -.10$, $p = .002$) trials, indicating that children who responded faster in the heart-only block tended to slow down more on the flower-only and mixed blocks.

As expected, accuracy on flower-only trials was positively related to academic ability (r = .18 to .30, all ps $< .001$) and prosocial behavior (r = .11, $p < .001$), and negatively related to conduct problems ($r = -.09$, $p < .01$) and hyperactivity ($r = -.20$, $p < .001$). Similarly, accuracy on mixed trials was positively related to academic ability (r = .25 to .38, all ps $< .001$) and prosocial behavior (r = .19, $p < .001$), and negatively related to conduct problems ($r = -.18$, $p < .001$) and hyperactivity ($r = -.26$, $p < .001$).

ΔRT for flower-only trials was negatively related to all academic outcomes ($r = -.06$ to $-.15$, all ps $< .05$) but was not significantly related to any behavioral outcome. ΔRT for mixed trials was not significantly correlated with any academic or behavioral outcomes.

Regression Models

Next, we present results from regression models predicting child academic and behavioral outcomes from mixed trial accuracy and RT, flower-only trial accuracy and RT, and the combination of the two trial types.

Mixed block. Results from the six mixed trial models, including F statistics, coefficients, and R^2 values, are presented in Table 2. Overall, mixed trial accuracy and ΔRT explained between 7% and 15% of the variance in academic outcomes, and between 2% and 7% of the variance in behavioral outcomes.

Academic outcomes. Accuracy on mixed trials was positively associated with all three academic outcomes (β = .23 to .36, $p < .001$), whereas ΔRT was negatively associated with all three outcomes ($\beta = -.09$ to $-.13$, $p < .05$). In other words, faster and more accurate children also had higher scores on all measures of academic ability.

In addition to these main effects, mixed trial accuracy and ΔRT interacted to predict the Applied Problems ($\beta = -.11$, $p < .01$) and Picture Vocabulary subscales ($\beta = -.10$, $p < .01$). This significant interaction means that the effect of RT on these two outcomes depends on the level of accuracy. Probing these interactions revealed that ΔRT became more predictive of child math and reading ability as accuracy increased (see Figure 1). For Applied Problems, the simple slope of ΔRT was two times greater when accuracy was high (75th percentile or 93% accuracy; $b = -.012$, $p < .001$) compared with when accuracy was low (25th percentile or 75% accuracy; $b = -.006$, $p = .004$). Regions of significance analyses revealed that the relationship between ΔRT and Applied Problems scores became significant when child accuracy was above 70%.

Similar results were found in the prediction of reading ability (e.g., Picture Vocabulary). The simple slope of ΔRT was 2 times greater when accuracy was high (75th percentile or 93% accuracy; $b = -.008$, $p = .003$) as opposed to when accuracy was low (25th percentile or 75% accuracy; $b = -.004$, $p = .04$). The relationship between ΔRT and PV scores became significant when child accuracy on mixed trials was above 74%.

It is important to note that among these models with significant interaction terms, the main effect of accuracy also remains significant. That is, at each value of ΔRT, higher accuracy is associated with higher math and reading ability.

Behavioral outcomes. Accuracy on mixed trials was associated in the expected direction with conduct problems ($\beta = -.16$, $p < .001$), hyperactivity ($\beta = -.27$, $p < .001$), and prosocial

Table 1
Descriptive Statistics and Bivariate Correlations for All Study Variables

Variable	1	2	3	4	5	6	7	8	9	10
1. Accuracy (mixed)	—									
2. ΔRT (mixed), ms	$.21^{***}$	—								
3. Accuracy (flower)	$.49^{***}$	$.16^{***}$	—							
4. ΔRT (flower), ms	$-.12^{***}$	$.47^{***}$	$-.15^{***}$	—						
5. Applied problems	$.38^{***}$	$-.01$	$.30^{***}$	$-.15^{***}$	—					
6. Letter-word	$.28^{***}$	$-.03$	$.18^{***}$	$-.07^{*}$	$.58^{***}$	—				
7. Picture vocabulary	$.25^{***}$	$-.03$	$.18^{***}$	$-.09^{**}$	$.52^{***}$	$.43^{***}$	—			
8. Conduct problems	$-.18^{***}$	$-.03$	$-.09^{**}$	$-.02$	$-.14^{***}$	$-.17^{***}$	$-.15^{***}$	—		
9. Hyperactivity	$-.26^{***}$	.03	$-.20^{***}$	.04	$-.29^{***}$	$-.29^{***}$	$-.21^{***}$	$.59^{***}$	—	
10. Prosocial behavior	$.19^{***}$	$-.04$	$.11^{***}$	$-.01$	$.19^{***}$	$.18^{***}$	$.17^{***}$	$-.68^{***}$	$-.56^{***}$	—
N	960	960	960	957	947	947	947	837	837	837
M	.81	547.42	.89	244.68	103.37	109.39	99.24	.26	.72	1.55
SD	.16	224.23	.17	206.09	13.66	12.17	9.92	.40	.64	.47

Note. RT = reaction time; ms = milliseconds.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 2
Direct and Interactive Effects of Mixed Trial Accuracy and ΔRT on Child Outcomes

Predictor	Applied problems	Letter-word	Picture vocabulary	Conduct problems	Hyperactivity	Prosocial behavior
Accuracy	.36***	.30***	.23***	−.16***	−.27***	.19***
ΔRT	−.13***	−.09*	−.11*	.01	.10**	−.08
Accuracy × ΔRT	−.11**		−.10**			
F(ndf, ddf)	$F(3, 651) = 38.27^{***}$	$F(2, 651) = 37.52^{***}$	$F(3, 651) = 19.00^{***}$	$F(2, 581) = 9.69^{***}$	$F(2, 581) = 26.78^{***}$	$F(2, 581) = 14.22^{***}$
Model R^2	.15	.09	.07	.02	.07	.04

Note. Models with nonsignificant interaction terms were reestimated to include only the direct effects of accuracy and reaction time (RT). ndf = numerator degrees of freedom; ddf = denominator degrees of freedom.
* $p < .05$. ** $p < .01$. *** $p < .001$.

behavior (β = .19, $p < .001$). On the other hand, mixed ΔRT only predicted hyperactivity (β = .10, $p < .01$) above and beyond accuracy. There were no interactive effects of mixed trial accuracy and ΔRT. Thus, as opposed to academic outcomes, mixed trial ΔRT was not as predictive of behavioral outcomes, and there was no evidence that ΔRT became more predictive at high levels of accuracy.

Flower-only block. Results from the six flower-only trial models, F statistics, coefficients, and R^2 values are presented in Table 3. Overall, accuracy and ΔRT on flower-only trials explained between 3% and 9% of the variance in academic outcomes, and only between 1% and 4% of the variance in behavioral outcomes.

Academic outcomes. Accuracy on flower-only trials was positively associated with all three academic outcomes (β = .15 to .27, $p < .001$), whereas ΔRT was negatively associated with scores on the Applied Problems subscale (β = −.13, $p < .01$). Thus, more accurate children had higher scores on math and reading ability, whereas faster children only showed higher scores on math ability. Accuracy and ΔRT also interacted to predict Applied Problems (β = −.11, $p < .01$). Similar to the findings for mixed trials, ΔRT on flower-only trials was more predictive of child math ability as accuracy increased (Figure 2a). Regions of significance analyses showed that the relationship between ΔRT and Applied Problems scores became significant when accuracy on flower-only trials was above 74%. The main effect of accuracy remained significant, meaning that higher accuracy was associated with better math ability, regardless of ΔRT.

Behavioral outcomes. Accuracy on flower-only trials was negatively associated with conduct problems (β = −.09, $p < .05$) and hyperactivity (β = −.17, $p < .001$), and positively associated with prosocial behavior (β = .10, $p < .05$). There were no main effects of ΔRT on any behavioral outcome. However, flower-only accuracy and ΔRT did interact to predict hyperactivity (β = .08, $p < .05$). Probing this interaction revealed that the relationship between ΔRT and hyperactivity became stronger as accuracy increased (Figure 2b). For example, ΔRT did not significantly predict hyperactivity when children were low (25th percentile or 82%; $b = .0002$, $p = .15$) or average (90%; $b = .0003$, $p = .05$) on accuracy, but it did predict hyperactivity when children were high on accuracy (75th percentile or 96%; $b = .0003$, $p = .03$). The relationship between ΔRT and hyperactivity became significant when flower-only accuracy was above 91%. The main effect of accuracy remained significant, meaning that higher accuracy was associated with less hyperactivity, regardless of ΔRT.

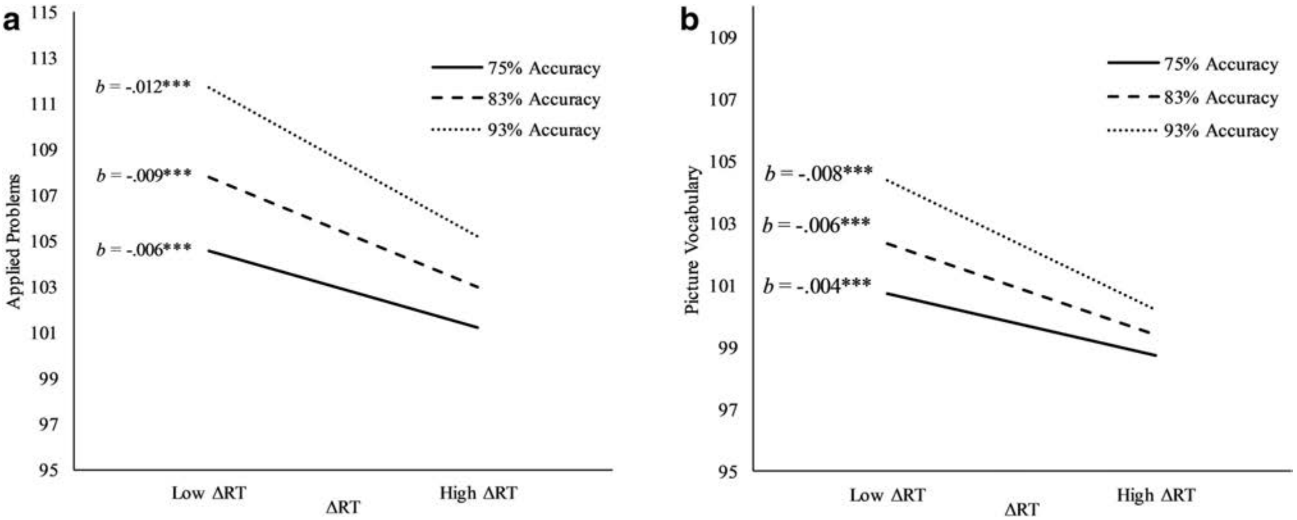

Figure 1. Interaction between mixed trial accuracy and RT (ΔRT) predicting (a) Applied Problems, and (b) Picture Vocabulary subtests. Values for low and high ΔRT represent the 10th and 90th percentiles, respectively.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3
Direct and Interactive Effects of Flower-Only Trial Accuracy and ΔRT on Child Outcomes

Predictor	Applied problems	Letter-word	Picture vocabulary	Conduct problems	Hyperactivity	Prosocial behavior
Accuracy	.27***	.15***	.15***	−.09*	−.17***	.10*
ΔRT	−.13**	−.03	−.04	−.02	.08	.00
Accuracy × ΔRT	−.11**				.08*	
F(ndf, ddf)	$F(3, 650) = 19.64^{***}$	$F(2, 650) = 8.04^{***}$	$F(2, 650) = 11.98^{***}$	$F(2, 581) = 3.14^{*}$	$F(3, 581) = 11.78^{***}$	$F(2, 581) = 2.97$
Model R^2	.09	.03	.03	.01	.04	.01

Note. Models with nonsignificant interaction terms were reestimated to include only the direct effects of accuracy and reaction time (RT). ndf = numerator degrees of freedom; ddf = denominator degrees of freedom.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Mixed and flower-only blocks. Our final six models simultaneously estimated the relationships between mixed and flower-only trial accuracy and ΔRT and child outcome measures. These findings are presented in Table 4. Accuracy and ΔRT on these two trial types jointly explained between 7% and 16% of the variance in academic outcomes, and between 3% and 8% of the variance in behavioral outcomes.

Academic outcomes. When mixed and flower-only scores were jointly included as predictors, mixed accuracy (β = .21 to .30, $p < .001$) and ΔRT (β = −.11 to −.13, $p < .01$) remained significant predictors of all three academic outcomes. The magnitude of effects and model R^2 values did not differ dramatically from the models that solely included mixed trials (see Table 2). In addition, the significant interactions between mixed accuracy and ΔRT predicting Applied Problems (β = −.08, $p < .05$) and Picture Vocabulary (β = −.09, $p < .05$) remained significant, though slightly reduced in magnitude.

Accounting for mixed trial performance, accuracy on flower-only trials only remained a significant predictor of Applied Problems scores (β = .12, $p < .05$). There was no longer a significant relationship between flower-only ΔRT and Applied Problems. The interaction of flower-only accuracy and ΔRT predicting AP was also no longer significant and was therefore trimmed from the model.

Behavioral outcomes. Similar to what we observed for academic outcomes, the significance and magnitude of effects for mixed accuracy and ΔRT predicting behavioral outcomes was largely unchanged compared with the mixed-only models. One exception is that mixed ΔRT significantly predicted prosocial behavior in this model (β = −.12, $p < .01$).

Accuracy on flower-only trials no longer predicted conduct problems and prosocial behavior in the joint model. The magnitude of the relationship between flower-only accuracy and hyperactivity also became smaller (β = −.10, $p < .05$). Whereas the direct effect of ΔRT remained nonsignificant for all behavioral outcomes, the interaction between flower-only accuracy and ΔRT predicting hyperactivity became insignificant in this joint model.

Discussion

The current study aimed to test whether accuracy and RT on the HF task, a common assessment tool used across wide age ranges, could be leveraged as joint indicators of child EF ability. We tested

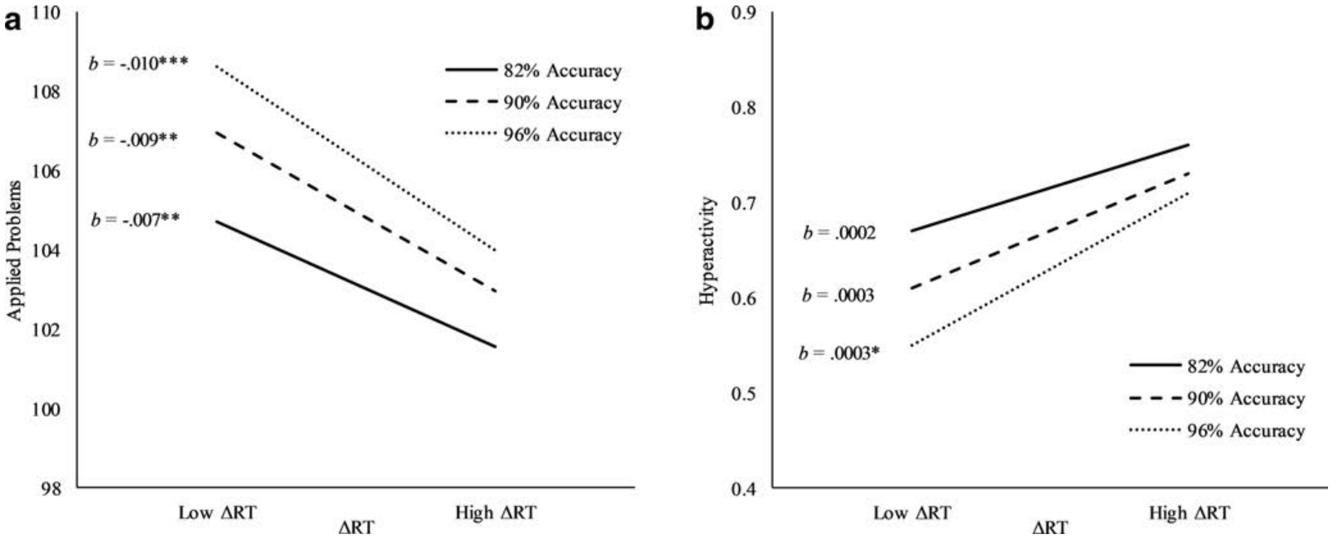

Figure 2. Interaction between flower-only trial accuracy and RT (ΔRT) predicting (a) applied problems, and (b) hyperactivity. Values for low and high ΔRT represent the 10th and 90th percentiles, respectively. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4
Direct and Interactive Effects of Mixed and Flower-Only Trial Accuracy and ΔRT on Child Outcomes

Predictor	Applied problems	Letter-word	Picture vocabulary	Conduct problems	Hyperactivity	Prosocial behavior
Accuracy (M)	.30***	.29***	.21***	−.16***	−.22***	.19***
ΔRT (M)	−.13**	−.11**	−.12**	.03	.12**	−.12**
Accuracy × ΔRT (M)	−.08*		−.09*			
Accuracy (F)	.12*	.04	.04	−.02	−.10*	.03
ΔRT (F)	.00	.04	.03	−.05	−.03	.07
Accuracy × ΔRT (F)						
F(ndf, ddf)	*F*(5, 650) = 24.14***	*F*(4, 650) = 19.46***	*F*(5, 650) = 12.27***	*F*(4, 581) = 5.16***	*F*(4, 581) = 15.29***	*F*(4, 581) = 7.71***
Model R^2	.16	.09	.07	.03	.08	.04

Note. Models with nonsignificant interaction terms were reestimated to include only the direct effects of accuracy and reaction time (ΔRT). M = mixed trials; F = flower-only trials; ndf = numerator degrees of freedom; ddf = denominator degrees of freedom.
* $p < .05$. ** $p < .01$. *** $p < .001$.

this question by modeling direct and interactive effects of accuracy and RT as predictors of six commonly studied outcome measures representing two broad domains of child development. Our findings indicate that even among early elementary-aged children, accuracy and RT interact in the prediction of child outcomes, with RT being a more informative index of EF ability for children who perform at high levels of accuracy. This pattern of findings remained significant in different task blocks (i.e., mixed, flower-only) and for different child outcome domains (i.e., academic, behavioral). Our finding of an interaction between these two metrics adds a layer of nuance to the existing EF assessment literature, which has so far tended to examine either accuracy or RT individually rather than yoking them together.

Among preschool and young school-age children, accuracy has been used as the primary metric of EF task performance, both in studies using the HF task (Blankson & Blair, 2016; Diamond et al., 2007) and other EF task batteries (e.g., Camerota, Willoughby, Kuhn, & Blair, 2018; Willoughby, Blair, & Family Life Project Investigators, 2016). In line with this, we found that HF accuracy had a direct effect on academic and behavioral outcomes in our sample of first-grade students. Therefore, there is support for the notion that accuracy is an informative index of EF ability among young children. Regarding RT, our data confirm that, on average, children do slow down when completing flower-only and mixed trials compared with heart-only trials. In some cases, this change in RT also had a direct effect on child outcomes, indicating that faster responding is informative above and beyond the contribution of accuracy. The weak correlation we observed between accuracy and RT on both task blocks further suggests that these two metrics represent different sources of information about child EF ability.

However, our finding of a significant interaction between accuracy and RT suggests that considering these two metrics in tandem may be more appropriate than considering them independently. Rather than operating as two independent indicators of EF ability, our findings indicate that the degree to which RT is an informative metric of EF performance depends on the child's accuracy. Specifically, at higher levels of accuracy, RT may be a more informative metric of child EF ability than at lower levels of accuracy. One way to interpret this finding is that if we were to consider two children with the same, high level of accuracy, the child who responds more quickly would be expected to demonstrate greater EF ability than the child who responds more slowly. However, RT may not play the same discriminatory role for two children who have the same, low level of accuracy. At low levels of accuracy, RT does not significantly predict child outcomes. On the other hand, the main effect of accuracy remained significant in all models, including those with significant interaction terms. This finding suggests that, across the board, higher accuracy on the HF task indicates greater EF ability among first graders.

Our finding that accuracy and RT interactively serve as an indication of child EF ability is noteworthy for several reasons. For one, the majority of studies using the HF task have not combined participant accuracy and RT in the same substantive models. In some cases (e.g., Diamond et al., 2007; Ursache & Raver, 2014), only RT or accuracy are considered, justified by the age group or mean level of performance of the study participants. That is, studies focused on young children or samples in which mean accuracy is relatively low tend to use accuracy scores, whereas studies with older children or samples in which mean accuracy is high tend to use RT scores. In other cases (e.g., Schonert-Reichl et al., 2015), both accuracy and RT are considered, but in separate models. By using both accuracy and RT in the same model, researchers can avoid making a priori or post hoc decisions about which measures are the best indicators of EF ability in a given sample. Additionally, they can more accurately model the performance of individuals within a sample who perform on the higher and lower ends of the spectrum.

For example, our findings suggested that RT became a meaningful predictor of academic outcomes when accuracy was around 75% or higher. These findings provide an empirically derived threshold that can be used to inform when RT might be a preferred metric over accuracy. However, even in a sample in which mean accuracy is above 75%, there may be a certain proportion of individuals scoring below that threshold. Therefore, adopting RT as the sole index of EF performance may misconstrue the EF ability of individuals scoring at lower levels of accuracy. Instead, incorporating accuracy and RT as independent and interactive predictors circumvents this problem by allowing for differential estimates of the effects of RT on substantive outcomes, based on each individual's observed accuracy.

As described in our analytic plan interpreting the regions of significance of the interactions in this study allows us to provide meaningful guidance about HF task scoring. Interestingly, our finding that RT becomes predictive when accuracy is above 75% is consistent with some efforts in the literature to combine accuracy and RT into single scores. For example, the two-vector scoring method for EF tasks in the NIH Toolbox (Zelazo et al., 2013) involves transforming accuracy and mean RT so that they are on the same scale (i.e., scores range from 0 to 5), and then adding these together in cases in which accuracy exceeds a threshold (i.e., 80%). In this case, accuracy and RT are treated as independent indices of EF performance, but only for individuals who achieve a high level of accuracy. Although it is unclear how the 80% threshold was chosen, the current findings suggest that there is some empirical grounding to the general approach of combining accuracy and RT given a certain level of performance. However, simply rescaling accuracy and RT such that they are on the same scale and adding them together implies that accuracy and RT are equally informative, which has yet to be empirically demonstrated.

Other approaches for combining accuracy and RT into single scores have been proposed, many of which stem from the speed–accuracy trade-off literature. According to this approach, individuals have control over the speed at which they complete a task, with the caveat that performing at faster speeds often sacrifices accuracy. Scoring methods that stem from the literature include the inverse efficiency score (IES), which is equal to the mean RT of correct responses divided by the percent of correct responses (Townsend & Ashby, 1978, 1983). Although this approach may be more parsimonious than using accuracy and RT as separate variables, some evidence suggests that the IES is not an ideal index of performance when accuracy is below 90% (Bruyer & Brysbaert, 2011), a criterion which is unlikely to be met in samples of young children. The drift diffusion model (Ratcliff & Rouder, 1998) uses correct and incorrect RT in an iterative distribution fitting approach, which results in the estimation of several parameters such as drift rate, boundary separation, and nondecision time. Particularly relevant here, boundary separation can best be thought of as an index of individuals' speed–accuracy trade-off "setting" (i.e., how "certain" a person must be before responding). Although this modeling technique jointly considers accuracy and RT, and has been applied to describe performance on EF tasks in children (e.g., Karalunas & Huang-Pollock, 2013), the substantive meaning of diffusion model parameters and their relation to child EF ability remains unclear. Therefore, although researchers are increasingly considering ways to combine accuracy and RT metrics in the scoring of EF tasks, there is not a single agreed-upon strategy about how to do so. Simpler methods, such as the two-vector scoring suggested by Zelazo and colleagues (2013), may oversimplify the contributions of accuracy and RT, whereas more complicated methods, such as IES and drift diffusion models, may be difficult to implement and interpret, given that they stem from diverse disciplines (e.g., psychometrics, cognitive psychology) and, in some cases, require the use of specialized analytic models (e.g., diffusion models).

The current approach, which uses accuracy and RT as interactive predictors of outcomes, is both intuitive and accessible to substantive researchers. First, it makes use of raw scores that are automatically generated from the HF task, without the need to apply complicated transformations or analytic models. Second, the approach can be implemented in a linear regression framework, which is analytically accessible to researchers in diverse fields. Third, it can be used flexibly with scores derived from different blocks of the HF task, which are hypothesized to represent different facets of EF ability (i.e., flower-only block as an index of inhibitory control, mixed block as an index of cognitive flexibility; Davidson et al., 2006). Finally, as demonstrated in the current article, we find that there are significant interactions between accuracy and RT in the prediction of a wide range of outcomes, and these interactions can be meaningfully interpreted.

Despite its strengths, a limitation of the current approach is that it is unknown how our conclusions would generalize to participants of different ages or in the prediction of different outcome measures. For example, if all participants in a study scored at a high level of accuracy, with little variability in observed accuracy scores, the interaction between accuracy and RT may not have the same magnitude or level of significance as observed here. We may similarly fail to find a significant interaction between accuracy and RT in cases in which all participants scored at a low level of accuracy (i.e., below 75%), suggesting that in certain situations, retaining individual scores may prove more useful than considering their joint contributions. However, this remains to be empirically tested. Additionally, although we selected six outcomes as exemplars of constructs that are common in developmental research, additional research is needed to determine whether these results generalize to other outcome measures. Although the current study tested concurrent criterion measures, another important next step may be to test how well these interactive models predict children's longer term outcomes. This type of examination may reveal whether children with slower RT and/or processing speed inevitably "catch up" to their faster peers. Finally, although the current study uses the HF task as one exemplar of a popular EF assessment, it is unclear whether our results would replicate with other widely used EF tasks (e.g., the DCCS task).

Another open question is how this approach (i.e., using accuracy and RT as joint indicators of EF performance) could be adapted when EF is the outcome, rather than the predictor, of interest.

Recently, psychometricians using factor analytic and item response theory techniques have proposed the use of a generalized linear modeling approach that makes use of item-level response and response time data as joint indicators of latent traits (Molenaar et al., 2015). In this approach, two latent variables are estimated to represent ability and speed. Whereas accuracy items load solely onto the ability factor, RT items load onto both factors, allowing for the parsing of RT variance into that which is indicative of ability and that which is indicative of speed. A recent article applied this approach to a variety of inhibitory control tasks administered with preschoolers and found that the joint use of accuracy and RT improved the precision of measurement of inhibitory control compared with a model that solely made use of accuracy (Magnus et al., 2017). It remains to be seen whether this type of approach could be used more widely, with children of different ages, and with tasks representing all three facets of EF ability.

In sum, the current study presents empirical support for using accuracy and RT measures in tandem as joint indicators of EF ability. We demonstrate that, even among young children, accuracy and RT on an EF task are both predictive of substantive outcomes, with RT being a stronger predictor for children who perform at a high level of accuracy. This intuitive, accessible approach represents one way to make use of the multiple indices of performance that result from a popular EF task, and represents an advancement in the current landscape of EF assessment, in which researchers tend to choose individual variables or groups of variables as indicators of performance, often lacking a clear theoretical rationale for these choices. It behooves other substantive researchers to investigate whether a similar approach could be applied to other EF tasks, to the prediction of other child outcomes, and among other age groups.

References

Bierman, K. L., Torres, M. M., Domitrovich, C. E., Welsh, J. A., & Gest, S. D. (2009). Behavioral and cognitive readiness for school: Cross-domain associations for children attending Head Start. *Social Development, 18,* 305–323. http://dx.doi.org/10.1111/j.1467-9507.2008.00490.x

Blair, C. (2002). School readiness. Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist, 57,* 111–127. http://dx.doi.org/10.1037/0003-066X.57.2.111

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78,* 647–663. http://dx.doi.org/10.1111/j.1467-8624.2007.01019.x

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.

Blankson, A. N., & Blair, C. (2016). Cognition and classroom quality as predictors of math achievement in the kindergarten year. *Learning and Instruction, 41,* 32–40. http://dx.doi.org/10.1016/j.learninstruc.2015.09.004

Brock, L. L., Rimm-Kaufman, S. E., Nathanson, L., & Grimm, K. J. (2009). The contributions of "hot" and "cool" executive function to children's academic achievement, learning-related behaviors, and engagement in kindergarten. *Early Childhood Research Quarterly, 24,* 337–349. http://dx.doi.org/10.1016/j.ecresq.2009.06.001

Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychologica Belgica, 51,* 5–13. http://dx.doi.org/10.5334/pb-51-1-5

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability. *Developmental Neuropsychology, 19,* 273–293.

Camerota, M., Willoughby, M. T., Kuhn, L. J., & Blair, C. B. (2018). The Childhood Executive Functioning Inventory (CHEXI): Factor structure, measurement invariance, and correlates in U.S. preschoolers. *Child Neuropsychology, 24,* 322–337. http://dx.doi.org/10.1080/09297049.2016.1247795

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. http://dx.doi.org/10.1016/j.neuropsychologia.2006.02.006

Davis, J. C., Marra, C. A., Najafzadeh, M., & Liu-Ambrose, T. (2010). The independent contribution of executive functions to health related quality of life in older women. *BMC Geriatrics, 10,* 16. http://dx.doi.org/10.1186/1471-2318-10-16

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

Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science, 318,* 1387–1388. http://dx.doi.org/10.1126/science.1151148

Diamond, A., & Kirkham, N. (2005). Not quite as grown-up as we like to think: Parallels between cognition in childhood and adulthood. *Psychological Science, 16,* 291–297. http://dx.doi.org/10.1111/j.0956-7976.2005.01530.x

Edgin, J. O., Mason, G. M., Allman, M. J., Capone, G. T., Deleon, I., Maslen, C., . . . Nadel, L. (2010). Development and validation of the Arizona Cognitive Test Battery for Down syndrome. *Journal of Neurodevelopmental Disorders, 2,* 149–164. http://dx.doi.org/10.1007/s11689-010-9054-3

Espy, K. A., McDiarmid, M. M., Cwik, M. F., Stalets, M. M., Hamby, A., & Senn, T. E. (2004). The contribution of executive functions to emergent mathematic skills in preschool children. *Developmental Neuropsychology, 26,* 465–486. http://dx.doi.org/10.1207/s15326942dn2601_6

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Child Psychology & Psychiatry & Allied Disciplines, 38,* 581–586. http://dx.doi.org/10.1111/j.1469-7610.1997.tb01545.x

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1,* 221–233.

Karalunas, S. L., & Huang-Pollock, C. L. (2013). Integrating impairments in reaction time and executive function using a diffusion model framework. *Journal of Abnormal Child Psychology, 41,* 837–850. http://dx.doi.org/10.1007/s10802-013-9715-2

Lakes, K. D., Bryars, T., Sirisinahal, S., Salim, N., Arastoo, S., Emmerson, N., . . . Kang, C. J. (2013). The Healthy for Life Taekwondo pilot study: A preliminary evaluation of effects on executive function and BMI, feasibility, and acceptability. *Mental Health and Physical Activity, 6,* 181–188. http://dx.doi.org/10.1016/j.mhpa.2013.07.002

Magnus, B. E., Willoughby, M. T., Blair, C. B., & Kuhn, L. J. (2017). Integrating item accuracy and reaction time to improve the measurement of inhibitory control abilities in early childhood. *Assessment*. Advance online publication. http://dx.doi.org/10.1177/1073191117740953

Miller, M., Nevado-Montenegro, A. J., & Hinshaw, S. P. (2012). Childhood executive function continues to predict outcomes in young adult females with and without childhood-diagnosed ADHD. *Journal of Abnormal Child Psychology, 40,* 657–668. http://dx.doi.org/10.1007/s10802-011-9599-y

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

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. http://dx.doi.org/10.1080/00273171.2014.962684

Pauli-Pott, U., & Becker, K. (2011). Neuropsychological basic deficits in preschoolers at risk for ADHD: A meta-analysis. *Clinical Psychology Review, 31,* 626–637. http://dx.doi.org/10.1016/j.cpr.2011.02.005

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356. http://dx.doi.org/10.1111/1467-9280.00067

Schonert-Reichl, K. A., Oberle, E., Lawlor, M. S., Abbott, D., Thomson, K., Oberlander, T. F., & Diamond, A. (2015). Enhancing cognitive and social-emotional development through a simple-to-administer mindfulness-based school program for elementary school children: A randomized controlled trial. *Developmental Psychology, 51,* 52–66. http://dx.doi.org/10.1037/a0038454

Séguin, J. R. (2004). Neurocognitive elements of antisocial behavior: Relevance of an orbitofrontal cortex account. *Brain and Cognition, 55,* 185–197. http://dx.doi.org/10.1016/S0278-2626(03)00273-2

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly, 22,* 173–187. http://dx.doi.org/10.1016/j.ecresq.2007.01.002

Snyder, H. R., Miyake, A., & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: Bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology, 6,* 328. http://dx.doi.org/10.3389/fpsyg.2015.00328

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), *Cognitive theory* (Vol. 3, pp. 200–239). Hillsdale, NJ: Erlbaum.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge, UK: Cambridge University Press.

Ursache, A., & Raver, C. C. (2014). Trait and state anxiety: Relations to executive functioning in an at-risk sample. *Cognition and Emotion, 28,* 845–855. http://dx.doi.org/10.1080/02699931.2013.855173

Vernon-Feagans, L., & Cox, M. (2013). I. Poverty, rurality, parenting, and risk: An introduction. *Monographs of the Society for Research in Child Development, 78,* 1–23. http://dx.doi.org/10.1111/mono.12047

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry, 57,* 1336–1346. http://dx.doi.org/10.1016/j.biopsych.2005.02.006

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. http://dx.doi.org/10.1037/pas0000152

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. http://dx.doi.org/10.1080/87565641.2010.549980

Woodcock, R. W., McGrew, K. S., & Mather, M. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities*. New York, NY: Riverside Publishing.

Wright, A., & Diamond, A. (2014). An effect of inhibitory load in children while keeping working memory load constant. *Frontiers in Psychology, 5,* 213. http://dx.doi.org/10.3389/fpsyg.2014.00213

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, 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,* 16–33. http://dx.doi.org/10.1111/mono.12032

Received January 23, 2019
Revision received March 11, 2019
Accepted March 20, 2019 ■