Journal of Applied Developmental Psychology 55 (2018) 4–13

Contents lists available at ScienceDirect

Journal of Applied Developmental Psychology

Assessing students' executive functions in the classroom: Validating a scalable group-based procedure

Jelena Obradović *, Michael J. Sulik, Jenna E. Finch, Nicole Tirado-Strayer

Stanford University, United States

ARTICLE INFO

Article history:
Received 8 October 2016
Received in revised form 10 February 2017
Accepted 9 March 2017
Available online 25 April 2017

Keywords:
Executive function
Assessment
Academic achievement
Classroom behaviors
Self-regulation

ABSTRACT

We describe and validate a novel, scalable, group-based assessment of executive functions (EFs) in a classroom setting using tablet computers. Relative to the conventional method of a more controlled, one-on-one individual assessment (IA), the group assessment (GA) can be administered quickly to many students, requires less training for assessors, and measures performance in a naturalistic classroom setting. In a socioeconomically and ethnically diverse sample of 269 students in third through fifth grade, we show that IA and GA scores for the same tasks were highly inter-correlated, equally reliable, and showed analogous associations with known EF covariates. IA and GA scores independently predicted teacher-rated self-regulated classroom behavior and standardized test scores. Further, only the GA score emerged as a unique predictor of academic achievement when controlling for prior achievement. We are sharing the tablet apps, source code, and supporting materials for this GA procedure at no cost under an open-source license.

1. Introduction

Executive function (EF) skills have been linked to various educational outcomes, including specific academic skills, school engagement, and self-regulated classroom behaviors (Diamond, 2013; Obradović, Portilla, & Boyce, 2012). However, the conventional approach to EF assessment is to measure children's performance on standard EF tasks in a highly controlled, laboratory-like setting, typically with a ratio of one child to one assessor. This approach lacks the ecological validity of assessment in a classroom setting—where children practice and apply EF skills daily—and does not scale well for collecting data from a large number of students. We developed a new procedure to simultaneously assess EF skills in all students in a classroom using standard EF tasks administered on tablet computers. The goals of the current study are to validate this new assessment by: (1) examining convergent validity with conventional individual assessment procedures; (2) comparing students' EF performance across group and individual assessment settings; (3) comparing associations of EFs with known demographic and educational covariates across the two assessment settings; and (4) investigating the predictive validity of EF skills assessed in group versus individual assessment settings for teachers' reports of students' self-regulated classroom behaviors and their academic achievement on standardized tests.

1.1. Executive functions and educational outcomes

EFs are a set of higher-order cognitive skills that enable children to inhibit their impulses, control inappropriate behaviors, ignore distractions, hold and manipulate information in the mind, and shift between competing rules or attentional demands. As such, EF skills are implicated in many aspects of school success. Over the last decade, researchers have linked direct assessments of EF skills to teachers' reports of students' self-regulated classroom behaviors, such as their ability to follow instructions, stay focused on tasks, and work collaboratively with peers (Ciairano, Visu-Petra, & Settanni, 2007; Diamond, 2013; Obradović et al., 2012; Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009). However, most of these studies have been conducted in early childhood. Researchers working with this age group often employ a composite of EF tasks tapping into multiple EF components (Fuhs, Farran, & Nesbitt, 2015; Neuenschwander, Röthlisberger, Cimeli, & Roebers, 2012; Sasser, Bierman, & Heinrichs, 2015). However, more research is needed to better understand how similar direct assessments of EFs relate to self-regulated classroom behaviors in middle childhood.

In addition to their role in promoting self-regulated behaviors, EF skills also contribute directly to academic performance. For example, solving math problems requires children to flexibly shift attention between different strategies and to manipulate and update information in working memory (Blair, Ursache, Greenberg, Vernon-Feagans, & Family Life Project Investigators, 2015). Although empirical evidence is most robust for the association between working memory and math skills (Bull & Lee, 2014; Jacob & Parkinson, 2015), meta-analytic studies have demonstrated that direct assessments of inhibitory control (Allan,

* Corresponding author.
E-mail address: jelena.obradovic@stanford.edu (J. Obradović).

http://dx.doi.org/10.1016/j.appdev.2017.03.003
0193-3973/

Hume, Allan, Farrington, & Lonigan, 2014), working memory (Friso-van den Bos, van der Ven, Kroesbergen, & van Luit, 2013) and cognitive flexibility (Yeniad, Malda, Mesman, van IJzendoorn, & Pieper, 2013) are all associated with children's performance on literacy and math achievement tests.

1.2. Ecological validity of executive function assessments

Researchers often assess children's EF skills in university-based laboratory settings using a battery of developmentally appropriate standard tasks administered by a highly trained research assistant (Carlson, 2005; Kochanska & Knaack, 2003). There are also many school-based studies, but these typically mimic a laboratory setting: researchers take children out of their classrooms to be assessed one-on-one in a quiet space such as a library room (Blair & Razza, 2007; Raver et al., 2011; Schmitt, McClelland, Tominey, & Acock, 2015; Weiland, Barata, & Yoshikawa, 2014). The assessor works closely with the child to explain the task instructions, provide guidance and feedback during practice trials, and ensure focused completion of the test trials. This context provides many external motivators (both intentional and inadvertent) for children to perform well on EF tasks — motivators that are not normally present in the classroom. Assessors are trained to establish good rapport with participants and express a caring and affirmative demeanor. They provide positive encouragement during practice trials, physical proximity during test trials, and praise after the task is completed. This individualized attention may motivate some children to (try harder to) perform well on the tasks and may contribute to artificially inflated EF performance that does not reflect the child's ability to engage EF skills in a more natural setting. Conversely, some children may be more comfortable in the classroom or better motivated by the presence of peers and teachers, and thus may underperform in a laboratory setting. Individual assessment minimizes the external distractions and interpersonal dynamics present in the classroom, and it provides controlled testing conditions that include constant monitoring and timed positive feedback (Silver, 2014), but it lacks ecological validity.

Ecological validity is an aspect of research design that refers to the similarity between the participants, materials, and settings used in a study and the real-world context under investigation (Shadish, Cook, & Campbell, 2002). By better aligning the assessment context with real-world conditions in which children employ their EF skills, researchers can improve the ecological validity of EF assessments (McCabe, Hernandez, Lara, & Brooks-Gunn, 2000; Sbordone, 2001). Specifically, assessing EF skills in a classroom setting, with its naturally occurring distractors and motivators, will yield a more ecologically valid measure of EF skills. It may also improve the predictive validity of directly assessed EF skills for students' self-regulated classroom behavior and measures of academic achievement such as performance on standardized tests.

1.3. Scalability of executive function assessments

As educators and policymakers debate the merits of assessing student progress using measures of socioemotional learning (Campbell et al., 2016; Duckworth & Yeager, 2015; Ursache, Blair, & Raver, 2012; West, 2016), researchers need to create valid, pragmatic, and cost-effective ways of assessing EFs at scale. Although teacher report on questionnaire measures of EF has been found consistently to predict children's academic achievement (Allan et al., 2014; McClelland, Acock, & Morrison, 2006), teacher report has several known limitations. First, teacher report of student behavior can be subject to a "halo" effect (Nisbett & Wilson, 1977), where the respondent's general impression of the child's overall functioning biases the report of specific skills. This may be exacerbated when teachers are required to rapidly evaluate and compare many students. There is also evidence of systematic racial and gender bias in teachers' reports (McKown & Weinstein, 2008; Ready & Wright, 2011). Moreover, when asked to consider students' self-regulation, teachers may find it difficult to differentiate between EFs and related constructs such as conscientiousness (Eisenberg, Duckworth, Spinrad, & Valiente, 2014). Further, questionnaire items tend to capture broad behavioral markers of self-regulation and composites tend to have positively skewed distributions, with many students scoring at or close to the scale maximum. As such, they are less sensitive than direct assessments in reflecting small differences in EF skills across students and incremental changes in EF skills over time. Finally, questionnaires require teachers to contribute considerable time and cognitive effort, which makes it difficult to gather information on all students in a classroom or to track changes throughout an academic year.

Direct assessment of EF skills addresses problems with objectivity and measurement precision (Silver, 2014) and is thus considered to be the "gold standard" of EF measurement. However, extant individual assessment procedures are prohibitively expensive for large-scale studies such as program evaluations. Further, taking children out of the classroom one at a time burdens teachers by reducing instructional time and disrupting students' attention and behavior. Understandably, teachers and district officials often object to this type of research design. In order to employ direct assessments of EF skills at scale, we need to develop a group-based assessment procedure that is pragmatic, cost-effective, and minimally disruptive.

Although researchers have recognized the need to measure EF skills in real-world settings (McCabe et al., 2000; Sbordone, 2001), they almost exclusively employ individual assessment procedures (Fuhs, Farran, & Nesbitt, 2013; Prager, Sera, & Carlson, 2016; Schmitt et al., 2015; Weiland et al., 2014). We were able to identify only one small pilot study (reported in a book chapter) in which EF data were collected in a group context. McCabe, Rebello-Britto, Hernandez, and Brooks-Gunn (2004) tested 44 preschoolers in a group administration procedure, where four familiar peers simultaneously completed modified laboratory-based tasks in a classroom setting with one administrator. The authors coded video recordings of group assessment and reported that children had a harder time controlling impulses during the Gift Wrap task when assessed in a peer group setting than during individual assessment, but otherwise did not compare children's EF performance across the two settings. Computerized tasks that automatically score accuracy and reaction time (RT) create an opportunity to extended this work and evaluate the feasibility of group assessment of EFs in middle childhood.

1.4. Current study

The main goal of the current study was to evaluate a new group assessment procedure that allows researchers to directly measure EF skills in all students at the same time. Our assessment procedures included a number of methodological innovations to obtain reliable and valid data while simultaneously reducing staff training requirements and the cost of data collection. We adapted developmentally appropriate, widely used EF tasks for administration on tablet computers. These tasks were selected to yield a broad measure of EFs, as represented by inhibitory control, working memory, and cognitive flexibility (see Measures for details). The computer-based tasks provided both accuracy and RT data, thus eliminating the need for video recording or coding of children's responses. Moreover, the portability of tablet devices and children's ease with the touch-screen interface enabled group assessment. Our procedure has the potential to significantly lower the costs and increase the widespread use of high quality direct assessments of EFs.

Our analyses compare the reliability and validity of this novel group assessment procedure with the reliability and validity of an analogous individual assessment procedure that was conducted in a quiet, highly controlled setting. We hypothesized that children's performance in the group assessment setting would show convergent validity with their performance in the more conventional individual assessment

setting. Further, we expected that EF skills would show similar associations with known demographic and educational covariates across both assessment settings. Finally, since direct assessment of EFs in the classroom has greater ecological validity, we hypothesized that performance in the group assessment setting would have greater predictive validity for children's self-regulated classroom behaviors and academic achievement compared to performance in the individual assessment setting.

2. Method

2.1. Participants

Students and teachers in 33 classrooms across eight schools in the San Francisco Bay Area participated in a three-tiered, longitudinal study design. First, all but one parent agreed to let their child participate in classroom activities, including the group assessment (GA) of EF skills. During GA, 720 students completed at least one of the three EF tasks evaluated in the current study. Second, we received parental written consent for 71% of these students to access their school records data. Third, to minimize burden on teacher time, we obtained teacher report of classroom behavior for a selected subsample of 334 students (approximately 10 per classroom). Separately, a subsample of 293 students was also tested in an individual assessment (IA) context for the purposes of the current study.

Thus, the primary analytic sample for this study includes 269 children (103 third-graders, 106 fourth-graders, and 60 fifth-graders; 52% female) with valid data for at least one EF task in both GA and IA settings. This sample was socioeconomically and ethnically diverse; among the 77% of parents who reported their children's ethnicity, children were identified as 7% African American, 23% Caucasian or White, 33% Asian or Pacific Islander, 32% Hispanic/Latino, and 5% multiracial or other. Among the 64% of parents who reported their educational attainment, 32% of parents had a high school education or less. This primary analytic sample ($N = 269$) was compared to children who completed the same EF tasks only in the GA setting ($N = 451$). There were no significant differences between these two samples in age, gender, ethnicity, or parent education. Children in the primary analysis sample had a slightly higher EF accuracy composite in the GA setting ($\beta = 0.13$, $t(716) = 2.08$, $p = .038$) than children who completed only the GA, but there was no significant group difference in the RT composite.

For analyses predicting academic achievement and teachers' reports of self-regulated classroom behavior, this sample was further restricted to the subset of participants for whom we had permission to access school records ($N = 204$) and for whom we collected teacher report ($N = 197$). Within the primary analytic sample ($N = 269$), there were no significant differences in EF performance, in both IA and GA contexts, between children who had school records data ($N = 204$) and those who did not ($N = 65$), as well as between children who had teacher report data ($N = 197$) and children for whom teacher report was not collected ($N = 72$).

The percent of missing data on EF variables ranged from 0% to 14%. Within the school records subsample, the percent of missing achievement data ranged from 5% to 19%. By study design, there was no missing data on teacher report of self-regulated classroom behavior. We used multiple imputation to generate 20 complete data sets for each of these three samples.

2.2. Procedures

Longitudinal data from four time points were used in this study. The participants described above (third- through fifth-graders) completed EF assessments in the fall of 2013 (Time 2), and teacher reports of students' classroom behavior were collected using an online questionnaire in the spring of 2014 (Time 3). Students' academic achievement data were obtained from school records for the previous academic year, 2012–13, when students attended second through fourth grades (Time 1), and for the following academic year, 2014–15, when students attended fourth through sixth grades (Time 4). Teachers were compensated with $95 for their participation and were given a tablet computer for classroom use.

2.2.1. Assessment of executive functions

Developmentally-appropriate and widely used EF tasks were adapted to be used on Android™ tablet computers. The tasks were designed to look fun and attractive to children by including cartoon pictures next to simplified task rules that students could easily read on their own. They were programmed to include a fixed number of practice trials and passcode-locked screens at designated intervals, to ensure that all students completed the tasks at a similar pace and were paying attention at the appropriate times to a research assistant who explained task instructions.

During GA, each child was given a tablet computer and all students in the classroom completed the EF tasks simultaneously. We developed a classroom procedure that enabled three research assistants (RAs) to administer EF assessment in a fashion similar to how teachers and their aides sometimes administer academic activities and tests. The lead RA, akin to a head teacher, stood at the front of the classroom, elicited students' attention, and explained the task rules using large posterboards that mimicked what students would later see on their tablet screens. The lead RA facilitated a group practice of each task rule by soliciting students' verbal responses. Afterward, all students had an opportunity to read the simplified task rules on their tablet screens and receive individualized feedback from the computer program, based on their performance on practice trials. The lead RA ensured that students completed practice and test trials at the same time by providing them with the passcodes to unlock different task blocks when all students were ready to proceed. Meanwhile, the other two RAs roamed the classroom like teacher's aides, helping students who needed technical assistance.

For IA, a subset of students completed the same tablet tasks outside of their classrooms in a quiet space such as a library room. A single RA explained the task rules to a child using a flipbook that contained the exact same images that were presented on posterboards in the GA context. Analogous to GA, the child was asked to verbally respond to two practice trials before proceeding to complete a practice block on a tablet computer. To account for practice effects, we administered IA in a counterbalanced fashion so that half of the students completed IA before GA and half of the students completed IA after GA. The amount of time between the IA and the GA ranged from 3 to 60 days ($M =$ 25.2 days; $SD = 17.6$ days).

2.3. Measures

2.3.1. Hearts and Flowers

The Hearts and Flowers (H&F) task, designed to assess inhibitory control and cognitive flexibility skills, has been widely used with elementary school students (Davidson, Amso, Anderson, & Diamond, 2006; Oberle & Schonert-Reichl, 2013; Roy, McCoy, & Raver, 2014; Yeniad et al., 2014). There were three blocks: (1) 12 congruent 'heart' trials, (2) 12 incongruent 'flower' trials, and (3) 33 mixed 'heart and flower' trials. Students were presented with an image of a red heart or flower on one side of the screen. For congruent heart trials, students were instructed to press the button on the same side as the presented stimuli (i.e., heart). For incongruent flower trials, students were instructed to press the button on the opposite side of the stimuli (i.e., flower). Accuracy scores were drawn from the incongruent block and the mixed block. Although the window of time in which children could respond (i.e., 750 ms) was based on previous research (Davidson et al., 2006), the pacing for the mixed block was too rapid for children in this study, resulting in many missing RT scores during

this block. Consequently, RT scores were drawn only from the incongruent block.

2.3.2. Multi-Source Interference Test

The Multi-Source Interference Test (MSIT) is a measure of inhibitory control skills that is used in middle childhood and adolescence (Bush & Shin, 2006; Liu, Angstadt, Taylor, & Fitzgerald, 2016; Ursache, Noble, & Blair, 2015). There were two blocks: (1) 24 congruent trials and (2) 24 incongruent trials. On both blocks, students were presented with a sequence of three digits. For each trial, two of these digits (the distractors) were the same and one (the target) differed from the distractors (e.g., "2 2 1"). Students were instructed to press a button whose numeric value corresponded to the numeric value of the target. For example, the correct response to the sequence "2 2 1" would be "1". For the congruent trials, the distractors were always zeroes and the numeric value of the target always corresponded to the numeric value of the correct button press (i.e., "1 0 0", "0 2 0", "0 0 3"). For the incongruent trials, the distractors were non-zero and the numeric value of correct button press was always *different* from the position of the correct response (e.g., "2 3 3", "2 2 1", "1 3 1"). Accuracy and RT scores from the incongruent block of trials were used.

2.3.3. Digit Span Backward

The Digit Span Backward (DSB) is a standard measure of working memory drawn from the Wechsler Intelligence Scale for Children-IV (Flanagan & Kaufman, 2009) that is commonly used in middle childhood (Blankenship & Bell, 2015; Brocki & Bohlin, 2004; St Clair-Thompson & Gathercole, 2006). A series of digits were presented sequentially on the tablet screen. The student was instructed to enter those numbers backwards onto a numeric keypad after the last digit was presented. There were four practice trials, each using strings that were two digits long. These practice trials were followed by eight test trials of increasing difficulty (two trials each of length two, three, four, and five digits). Accuracy scores were computed as the proportion of correct test trials.

2.3.4. Scoring of EF tasks

Anticipatory responses—defined as a response <200 milliseconds (ms) after stimulus presentation—were recoded as missing for the accuracy scores and RT scores. Further, the H&F and MSIT tasks were timed, such that students were unable to respond after 750 ms and 2500 ms, respectively. If the student failed to respond during this window, the trial was counted as incorrect for the accuracy score and as missing for the RT score. Finally, as is standard practice, RT scores were calculated only for the accurate trials and were not calculated for the first trial in each block. To receive an accuracy score for each task, a participants were required to have non-missing scores for 8/8 DSB trials (because difficulty varied across trials), 5/12 H&F incongruent trials, 10/33 H&F mixed trials, and 15/45 MSIT incongruent trials. In addition, RT scores were based on a *minimum* of three accurate trials for the H&F incongruent block and the MSIT incongruent block. Three outliers, defined as accuracy or RT scores that were >4 *SD* above or below the sample mean, were Winsorized to the highest non-outlier value that was observed for that task.

Cronbach's alpha reliability for the RT variables was as follows: MSIT incongruent IA = 0.91, GA = 0.90; H&F incongruent block IA = 0.88, GA = 0.90. Since Cronbach's alpha systematically underestimates reliability of scale using binary indicators (Raykov, Dimitrov, & Asparouhov, 2010), we used tetrachoric correlations to correct for this bias in our computation of alpha coefficient for the binary accuracy variables. Alpha reliability for the accuracy variables was as follows: digit span IA = 0.85, GA = 0.82; MSIT incongruent block IA = 0.90, GA = 0.95; H&F incongruent block IA = 0.91, GA = 0.89; H&F mixed block IA = 0.92, GA = 0.91.

2.3.5. Academic achievement

State-administered standardized test scores were used to measure English/language arts and mathematics skills at two time points. In spring of 2013, scores were drawn from the California Standards Test (California Department of Education, 2016a), a test designed to match the state's academic content standards. In spring of 2015, scores were drawn from the Smarter Balance Assessment Consortium (California Department of Education, 2016b), a test designed to match the new Common Core State Standards.

2.3.6. Self-regulated classroom behaviors

Teachers reported on students' classroom behaviors relevant to academic success using the Teacher-Child Rating Scale (TCRS; Hightower, 1986). Each item on the *task orientation* (e.g., "well-organized", "completes work", "works well without adult support") and *frustration tolerance* (e.g., "accepts things not going his/her way", "copes well with failure") scales was rated on a five-point scale, ranging from 1 = "not at all" to 5 = "very well". The reliability of these two 5-item scales was high (αs = 0.88 and 0.91). These scales were highly correlated (r = 0.66) and were averaged to create a score assessing self-regulated classroom behaviors.

2.3.7. Covariates

Child age in years, child gender (0 = male, 1 = female), and years of parent education were included as covariates.

2.4. Analysis plan

We performed a series of four analyses to validate GA and describe similarities and differences between IA and GA. First, we examined descriptive statistics and the correlations between IA and GA for each task. Second, we compared associations of EF composites with known demographic and educational covariates across the two assessment settings. Third, we examined whether there were mean differences between IA and GA performance composites while accounting for the ordering of the two assessments. Finally, we tested the unique contribution of EF performance in IA and GA settings for the prediction of self-regulated classroom behaviors and academic achievement.

3. Results

3.1. Descriptive statistics and correlations

3.1.1. Individual executive function tasks

Correlations and descriptive statistics for the scores on the individual EF tasks are presented in Table 1. Only accuracy for the incongruent blocks on MSIT and H&F tasks showed an indication of ceiling effects. For the MSIT incongruent block, 20% of students attained perfect scores during IA, and 19% of students attained perfect scores during GA. For the H&F incongruent block, 38% of students attained perfect scores during IA, and 34% of students attained perfect scores during GA. Although extant studies do not explicitly state how many children hit ceiling, they report comparable mean and standard deviation statistics on these tasks (Liu et al., 2016; Oberle & Schonert-Reichl, 2013; Yeniad et al., 2014).

The accuracy scores across tasks were generally moderately correlated, with rs ranging from 0.13 to 0.52 for IA and from 0.16 to 0.42 for GA. Similarly, the RT scores across tasks were positively related for IA (r = 0.44, $p < 0.001$) and for GA (r = 0.33, $p < 0.001$). Consistent with prior research, the accuracy and RT scores were negatively correlated within each task block (IA H&F incongruent block: $r = -0.55$, $p < 0.001$; GA H&F incongruent block: $r = -0.49$, $p < 0.001$; IA MSIT incongruent block: $r = -0.28$, $p < 0.001$; and GA MSIT incongruent block: $r = -0.23$, $p < 0.001$).

Table 1
Descriptive statistics and correlations among executive function individual task scores.

		1	2	3	4	5	6	7	8	9	10	11	12
1	IA DSB Acc	–											
2	IA H&F Inc Acc	0.34***	–										
3	IA H&F Mix Acc	0.33***	0.52***	–									
4	IA MSIT Inc Acc	0.32***	0.13*	0.22***	–								
5	GA DSB Acc	0.44***	0.29***	0.37***	0.32***	–							
6	GA H&F Inc Acc	0.19**	0.31***	0.29***	0.11	0.28***	–						
7	GA H&F Mix Acc	0.15*	0.32***	0.52***	0.18**	0.34***	0.42***	–					
8	GA MSIT Inc Acc	0.30***	0.14*	0.12	0.48***	0.38***	0.21***	0.16**	–				
9	IA H&F Inc RT	−0.23***	−0.55***	−0.57***	−0.09	−0.24***	−0.33***	−0.34***	−0.08	–			
10	IA MSIT Inc RT	−0.34***	−0.42***	−0.53***	−0.28***	−0.28***	−0.19**	−0.21***	−0.23***	0.44***	–		
11	GA H&F Inc RT	−0.05	−0.26***	−0.27***	0.01	−0.15*	−0.49***	−0.52***	−0.07	0.43***	0.16**	–	
12	GA MSIT Inc RT	−0.19**	−0.17**	−0.14*	−0.14*	−0.25***	−0.31***	−0.39***	−0.23***	0.13*	0.38***	0.33***	–
	Mean	0.51	0.84	0.48	0.87	0.46	0.86	0.51	0.85	561	1400	544	1383
	SD	0.20	0.18	0.19	0.13	0.20	0.15	0.20	0.17	68	228	73	225
	Min	0.00	0.25	0.12	0.33	0.00	0.33	0.06	0.08	385	860	383	672
	Max	1.00	1.00	0.94	1.00	1.00	1.00	0.97	1.00	718	2014	734	2100
	N	262	245	232	253	252	266	248	268	244	253	266	267

Note. IA = individual assessment, GA = group assessment; DSB = Digit Span Backwards, H&F = Hearts and Flowers; MSIT = Multi-Source Interference Test. Inc = incongruent block; Mix = mixed block. Acc = accuracy; RT = reaction time. Correlations, *M*s, and *SD*s use multiply imputed data ($N = 269$).
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

3.1.2. Executive function composites

We aggregated individual EF task accuracy and RT variables to create more reliable measures of EF skills within IA and within GA. The DSB, H&F incongruent block, H&F mixed block, and the MSIT incongruent block accuracy scores were standardized and averaged to create accuracy composite scores (IA: $\alpha = 0.65$; GA: $\alpha = 0.63$). Similarly, the RT variables for the H&F incongruent block and the MSIT incongruent block were standardized and averaged to create RT composite scores.

Correlations among the EF IA and GA composite scores and other study variables are presented in Table 2. The IA and GA accuracy composites were strongly positively correlated ($r = 0.59$, $p < 0.001$); similarly, the IA and GA RT composites were positively correlated ($r = 0.40$, $p < 0.001$). The accuracy and RT composites were strongly negatively correlated within the IA ($r = -0.64$, $p < 0.001$) and within the GA ($r = -0.54$, $p < 0.001$). Relative to boys, girls performed modestly better for the IA accuracy ($r = -0.15$, $p < 0.05$) and GA accuracy ($r = -0.17$, $p < 0.01$) scores and faster for the IA RT score ($r = 0.19$, $p < 0.01$). Girls were also rated as higher in self-regulated classroom behaviors by their teachers ($r = 0.25$, $p < 0.001$). Older children had greater IA accuracy ($r = 0.31$, $p < 0.001$) and GA accuracy ($r = 0.44$, $p < 0.001$), as well as faster IA RT ($r = -0.30$, $p < 0.001$) and GA RT ($r = -0.35$, $p < 0.001$). Finally, the EF accuracy and RT composites across both assessment contexts were moderately to strongly correlated with the academic achievement variables and teachers' reports of self-regulated classroom behaviors. As expected, these correlations were positive for accuracy and negative for RT variables. We used Fisher's *r*-to-*z* transformation to test whether the correlations between the EF composites and main covariates and outcome variables differed across IA and GA settings; there were no significant differences.

Table 2
Bivariate correlations and descriptive statistics for all study variables.

		1	2	3	4	5	6	7	8	9	10	11	12
1	IA Acc comp (T2)	–											
2	GA Acc comp (T2)	0.59***	–										
3	IA RT comp (T2)	−0.64***	−0.41***	–									
4	GA RT comp (T2)	−0.27***	−0.54***	0.40***	–								
5	Self-reg behavior (T3)	0.29***	0.32***	−0.15*	−0.22***	–							
6	ELA (T1)	0.43***	0.40***	−0.20***	−0.14*	0.57***	–						
7	Math (T1)	0.46***	0.35***	−0.32***	−0.14*	0.43***	0.74***	–					
8	ELA (T4)	0.49***	0.50***	−0.27***	−0.26***	0.55***	0.78***	0.63***	–				
9	Math (T4)	0.53***	0.50***	−0.37***	−0.25***	0.49***	0.71***	0.74***	0.80***	–			
10	Female student	−0.15*	−0.17**	0.19**	0.11	0.25***	0.12	−0.07	0.11	−0.04	–		
11	Student age	0.31***	0.44***	−0.30***	−0.35***	−0.07	0.02	−0.01	0.20**	0.15*	−0.19**	–	
12	Parental education	0.16*	0.12	−0.06	−0.03	0.11	0.32***	0.29***	0.39***	0.34***	0.03	0.04	–
	Mean	0.00	0.00	0.00	0.00	3.59	362	405	2500	2513	0.52	9.91	13.76
	SD	1.00	1.00	1.00	1.00	0.93	73	97	97	93	0.50	0.87	3.48
	Min	−2.50	−3.11	−2.02	−2.33	1.20	187	191	2266	2243	0	8.00	8.00
	Max	1.49	1.36	2.27	1.97	5.00	600	600	2701	2728	1	12.29	18.00
	% missing	0%[a]	0%[a]	0%[a]	0%[a]	1%[b]	6%[c]	5%[c]	19%[c]	19%[c]	0%[c]	0%[c]	16%[c]

Note. IA = individual assessment; GA = group assessment; Acc comp = accuracy composite; RT comp = reaction time composite. T1 = Time 1 (Spring 2013); T2 = Time 2 (Fall 2013); T3 = Time 3 (Spring 2014); Time 4 (Spring 2015). Self-reg behavior = self-regulated classroom behaviors, as indexed by teacher report of task orientation and frustration tolerance. ELA = English/Language Arts.
[a] $N = 269$ for EF variables.
[b] $N = 199$ for teacher report of students' self-regulated behaviors.
[c] $N = 206$ for achievement and demographics variables drawn from school records data. Correlations, *M*s, and *SD*s use multiply imputed data ($N = 269$).
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

3.2. Mean differences

To test whether there were mean differences in performance between IA and GA scores (0 = GA; 1 = IA), we estimated a multilevel model for each accuracy and RT variable in which scores were nested within individuals and individuals were nested within classrooms. We included school fixed effects as covariates to control for any school-level differences in students EFs and achievement. Including school fixed effects removes any biases from omitted variables that are constant within schools and accounts for the nesting of classrooms within schools (Allison, 2009). It is a recommended modeling strategy when the number of schools is relatively small and not the focus of the analyses (McNeish & Wentzel, 2016). In addition, covariates included age, gender, ethnicity, parent education, and a dummy variable representing the assessment order effect (i.e., whether the assessment was done first: 0 = No; 1 = Yes), given the counterbalanced IA/GA assessment design. Results for these models are presented in Table 3.

There were assessment order effects for all six EF variables. For all four accuracy scores, children performed better the second time they were assessed: *b*s ranged from −0.04 to −0.11 and correspond to the difference in the accuracy proportion for the first assessment relative to the second assessment. There was also evidence of assessment order effects for RT. On average, RT was 36 ms faster for the H&F incongruent block and 171 ms faster for the MSIT incongruent block when children were assessed the second time (relative to the first time).

Some differences in average performance between IA and GA emerged; however, the direction of these effects was not entirely consistent. Children's average accuracy on the MSIT incongruent block ($b = 0.02$, $t(240.66) = 2.50$, $p = 0.013$) and DSB task ($b = 0.05$, $t(236.28) = 4.05$, $p < 0.001$) was greater in IA compared to GA. In contrast, children's average RT during the H&F incongruent block was slower in IA relative to GA ($b = 12.93$, $t(237.56) = 3.09$, $p = 0.002$). There were no significant differences between GA and IA for H&F incongruent and mixed block accuracy and for MSIT incongruent block RT. We also tested interactions between task order and type of assessment setting. The interaction was significant only for DSB ($b = -0.10$, $t(241.53) = -2.67$, $p = 0.008$). For children who were being assessed the first time, there was no difference in performance between IA and GA. However, for children who were being assessed the second time, performance was better for IA than for GA ($b = 0.10$, $t(245.51) = 4.58$, $p < 0.001$).

3.3. Predictive validity

We estimated a series of multilevel models to test whether teachers' reports of students' self-regulated classroom behaviors in the spring of the focal assessment year (T3) and academic achievement on state tests the following year (T4) were independently predicted by the IA and GA EF composite scores. First, we examined whether accuracy and RT composite scores measured in each assessment context were predictive of each dependent variable. Then, we examined the IA and GA scores within a single model. In all of these models, we accounted for the nesting of children within classrooms and included school fixed effects, child age, gender, and ethnicity, and parent education as covariates. For the achievement variables, we ran the models with and without controlling for prior academic achievement at T1. Teachers' reports of student behavior were collected on only one occasion, so it was not possible to control for earlier teachers' reports of self-regulated classroom behaviors. When IA accuracy and RT were tested as predictors and when GA accuracy and RT were tested as predictors, only the accuracy variables were significantly related to the dependent variables. Based on the null results for RT, we only included the accuracy composites in our subsequent models testing the joint contribution of IA and GA to self-regulated classroom behaviors and academic achievement.

Results for these models are presented in Table 4. IA accuracy and GA accuracy were each independently predictive of teachers' reports of self-regulated classroom behavior (*b*s = 0.23 and 0.51, $t(204.89) = 2.26$ and $t(201.09) = 4.85$, $p = 0.025$ and $p < 0.001$), as well as ELA achievement (*b*s = 27.97 and 41.14, $t(185.77) = 3.17$ and $t(222.36) = 4.55$, *p*s < 0.001) and math achievement (*b*s = 30.70 and 34.96, $t(186.91) = t(199.84) = 3.57$ and 3.89, *p*s < 0.001) when prior achievement was not included as a predictor. Controlling for prior academic achievement (see Model 2), IA accuracy was no longer related to ELA or math. However, GA was still significantly related to both ELA ($b = 19.64$, $t(179.90) = 2.74$, $p = 0.007$) and math ($b = 22.73$, $t(126.26) = 2.92$, $p = 0.004$).

We conducted follow-up analyses to examine whether the interaction between assessment order and the IA and GA accuracy scores predicted school outcomes. We found no significant interactions between assessment order and EF scores when predicting math or ELA

Table 3
Mean differences between EF performance in individual and group assessment contexts.

	Accuracy		Reaction time	
	b	(*se*)	*b*	(*se*)
Hearts and Flowers, incongruent block				
Intercept	0.914	(0.064)***	499.694	(25.840)***
First assessment	−0.063	(0.012)***	36.296	(4.194)***
Context (0 = GA, 1 = IA)	−0.014	(0.012)	12.933	(4.188)**
Hearts and Flowers, mixed block				
Intercept	0.510	(0.067)***		
First assessment	−0.111	(0.010)***		
Context (0 = GA, 1 = IA)	−0.019	(0.010)		
MSIT, incongruent block				
Intercept	0.867	(0.063)***	1394.060	(82.646)***
First assessment	−0.041	(0.010)***	170.760	(11.634)***
Context (0 = GA, 1 = IA)	0.024	(0.010)*	0.362	(11.613)
Digit Span Backwards				
Intercept	0.437	(0.066)***		
First assessment	−0.035	(0.013)**		
Context (0 = GA, 1 = IA)	0.054	(0.013)***		

Note. School fixed effects and demographic covariates (child's age, gender, ethnicity, and parent education) were included in the models.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

Table 4
Multilevel regression analyses predicting students' self-regulated classroom behavior and academic achievement.

	Self-regulated classroom behaviors (T2)	ELA achievement (T3)		Math achievement (T3)	
	Model 1 *b* (*se*)	Model 1 *b* (*se*)	Model 2 *b* (*se*)	Model 1 *b* (*se*)	Model 2 *b* (*se*)
IA EF accuracy composite (T2)	0.233* (−0.103)	27.971** (−8.815)	5.533 (−6.969)	30.696*** (−8.591)	5.032 (−7.639)
GA EF accuracy composite (T2)	0.512*** (−0.106)	41.140*** (−9.048)	19.641** (−7.173)	34.959*** (−8.989)	22.731** (−7.792)
Prior ELA/math (T1)			0.796*** −0.067		0.552*** −0.056

Note. IA = individual assessment; GA = group assessment. ELA = English/Language Arts. T1 = Time 1 (Spring 2013); T2 = Time 2 (Fall 2013); T3 = Time 3 (Spring 2014); Time 4 (Spring 2015). School fixed effects and demographic covariates (child's age, gender, ethnicity, and parent education) were included in the models.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

achievement. This was true both for models in which we did not control for prior test scores and for models where we did control for prior test scores. There was an interaction between assessment order and IA accuracy (but not between assessment order and GA accuracy) for teachers' reports of self-regulated classroom behaviors. For children who completed IA after GA, IA accuracy was positively related to children's self-regulated classroom behaviors, $b = 0.73$, $t(244.39) = 4.81$, $p < 0.001$. For children who completed IA before GA, IA accuracy was unrelated to children's self-regulated classroom behaviors, $b = 0.01$, $t(140.10) = 0.07$, $p = 0.943$.

4. Discussion

In the current study, we demonstrated that a novel, group-based approach to measuring students' EF skills in a classroom context is a reliable and valid alternative to a conventional, individual assessment approach. One important advantage of the group-based assessment procedure over the conventional individual assessment approach is that it provides a way to simultaneously collect direct tests of EFs for all students in a class, making this a pragmatic and scalable method of data collection for school-based studies. Further, the group assessment procedure, designed to improve the ecological validity of EF measurement, yielded a measure of EF skills that uniquely predicted teacher report of self-regulated classroom behaviors and solely predicted a two-year improvement in ELA and math achievement scores.

Despite the established importance of EF skills for school success—via self-regulated classroom behaviors that support learning as well as specific academic skills (Ciairano et al., 2007; Rimm-Kaufman et al., 2009)—extant school-based studies assess student EFs in quiet, controlled settings that minimize the distractions that are present in a classroom environment (Blair & Razza, 2007; Fuhs et al., 2013; Raver et al., 2011; Schmitt et al., 2015). We aimed to demonstrate that standard EF tasks can be employed in a more naturalistic classroom setting, directly assessing EF skills for all students at the same time. We adapted standard EF tasks to run as apps on tablet computers and developed a scripted classroom procedure that enabled three research assistants to administer the assessment, akin to how lead teachers and their aides conduct academic activities. This involved minor EF task modifications, such as incorporating more practice trials and using passcode-locked screens to control the pacing of the assessment and enable an RA to deliver task instructions to the full group. An important advantage of this tablet-based assessment method is that it requires significantly less time and training for personnel relative to data collection using traditional "table-top" EF assessment procedures. It also obviates the need for manual task scoring because the tablet computers automatically record accuracy and RT of students' responses.

Our findings show evidence of convergent validity in that student EF performance in the classroom setting was strongly correlated with performance on the same tasks administered in a quieter, individual assessment outside the classroom. This was true for measures of accuracy and for RT across all three EF tasks. Individual EF tasks' accuracy and RT variables showed similar internal reliability of task trials across two assessment settings. Moreover, the accuracy composite variables, which reflected overall student EF performance, showed almost identical internal reliability across the two assessment settings. Similarly, the strength of correlation between the two RT variables did not vary across settings. Finally, the bivariate associations of EF accuracy composites with student demographic variables, teacher report of self-regulated classroom behavior, and standardized measures of academic achievement did not differ across the two settings. Together, these results demonstrate that direct assessment of EF skills in a group setting is as reliable and valid as direct assessment of EF skills in a conventional, individual setting. As such, group-based assessment is a viable option for researchers interested in using direct tests of EF skills in large-scale program evaluation studies or in school- or district-wide assessments.

Although we compared students' average performance across the two settings, we did not expect that mean-level scores would be clearly better in one setting, because sources of motivation and distraction differ in each context. For example, some students may be motivated to perform better in individual assessment due to the caring demeanor, individualized feedback, and physical proximity of the assessor, whereas other students may be motivated to perform better in the group setting due to interpersonal dynamics with peers and teachers. Indeed, we found small differences in average performance across the two settings, going in both directions. On average, students performed better in the group than in the individual assessment setting on a measure of inhibitory control (Hearts & Flowers task), as indexed by slightly lower RTs on accurate trials. In contrast, students performed better in the individual than in the group setting on tasks designed to assess their ability to inhibit interference (Multi-Source Interference Test) and engage working memory (Digit Span Backwards task). We can only speculate about the reasons for these differences. Perhaps the mere presence of peers may exert pressure on students to respond faster when suppressing a dominant motoric response, whereas classroom distractions may hinder attention focusing or manipulation of mental information. Future studies should investigate contextual factors that may explain these mean-level performance differences and identify which types of students are more likely to perform better in a given context.

By design, group-based EF assessment has greater ecological validity than traditional individual EF assessment, as it is more similar to the real-world conditions in which students rehearse and apply EF skills (Shadish et al., 2002). Unlike external validity, ecological validity is not a testable construct; however, we hypothesized that the greater ecological validity of the group assessment setting would improve predictive validity of EF scores. Our study revealed that the accuracy of student EF performance in both assessment settings independently predicted teachers' perceptions of self-regulated classroom behavior. Specifically, the two analogous measures of EF skills were uniquely related to teacher report of students' ability to be organized, complete classwork, ignore distractions, work independently, accept imposed limits, cope with failure, and tolerate frustration. Likewise, the two accuracy composites independently predicted student performance on standardized tests of ELA and math achievement the following year. In contrast, accuracy on EF tasks in the group setting was the sole EF predictor of both math and ELA achievement test scores across two academic years when controlling for prior achievement. The unique contribution of EF skills assessed in the group context for the longitudinal change in academic skills is notable, given the robust longitudinal stability of achievement test scores (Duncan et al., 2007). In our study, repeated measures of academic achievement shared more than half of their variance.

We propose that EF skills assessed in an individual setting capture the child's EF capacity in a controlled testing environment, whereas EF skills assessed in a group context capture the child's ability to engage the same EF skills in a naturalistic setting characterized by distractions and interpersonal dynamics. Our findings demonstrated that EF skills measured in both contexts are relevant for students' ability to control their attention, behaviors, and emotions in the classroom, as well as for their performance on standardized academic tests. However, only EF performance in a classroom environment emerged as a significant predictor of changes in students' achievement test scores, empirically corroborating the notion that improved ecological validity of classroom assessment may also improve predictive validity of directly assessed EF skills for school success.

Investigators typically use accuracy scores to measure EF in early childhood and early school-age children, whereas RT scores are typically used in studies of adolescents and adults. Accuracy scores are sensitive to large between-person differences in EFs in younger children, but as children get older these scores approach a ceiling for many EF measures and eventually they no longer vary meaningfully across persons. However, RTs can be used to assess EFs in older participants, even when using comparatively simple EF tasks that are appropriate for children (Best & Miller, 2010). In this study, we measured accuracy and RT for two EF

tasks to gain a more complete view of children's EF during middle childhood, a developmental period that has been studied less thoroughly than early childhood (Hughes, 2011). There are few general guidelines about the age at which it is appropriate to switch from using accuracy to RT for EF measurement. This could be because even superficially similar EF tasks can differ substantially in difficulty (Lagattuta, Sayfan, & Monsour, 2011) and because performance differs depending on socioeconomic status (Vernon-Feagans, Cox, & The Family Life Project Key Investigators, 2013), making general recommendations difficult.

In this study, two EF RT variables showed excellent internal reliability and expected associations with corresponding EF accuracy variables and with performance on other EF tasks. Further, they were strongly negatively correlated with each other, and each was correlated in expected ways with children's age, academic achievement, and self-regulated classroom behavior. However, when both accuracy and RT composites were included in the same model, only the accuracy composite was uniquely associated with ELA scores, math scores, and teachers' reports of self-regulated classroom behaviors. These results suggest that the switch from accuracy to RT measures of EF (at least for the H&F and MSIT tasks) should occur after the age period studied here (third through fifth grade), although lower reliability for the RT composite (due to fewer tasks being represented by the RT composite than the accuracy composite) could also have contributed to weaker predictive findings for RT. In any case, our tablet assessment provides both scores, facilitating continuity of assessment across time and across a range of participants' ages.

4.1. Limitations and future directions

There are several important limitations to our study. First, although the same child completed both IA and GA (which was necessary in order to examine how these two different assessments were related), the order of these assessments could potentially influence the scores for the later assessment (e.g., due to practice effects). To mitigate this issue, we counterbalanced the assessment order, which allowed us to investigate whether assessment order was systematically related to performance: indeed, performance was consistently better on the later assessment. A second limitation is that the response window for the mixed Heart & Flower block trials was too short for children in this study. Future studies should conduct thorough pilot testing to ensure appropriate specification of each task that will be used in a field setting. A third limitation was that we did not measure systematic differences between classrooms that could affect the testing environment. We did account for the nesting of children within classrooms and we controlled for school fixed effects. Future research should investigate whether GA is a better predictor of academic achievement than IA among children who regularly experience a more distracting learning environment that places strong demands on attentional control and other EF skills. In such cases, we posit that classroom-based assessment would better reflect the real-world application of EF skills. A fourth limitation was that, unlike the longitudinal measures of academic achievement, teachers' reports of self-regulated classroom behaviors was collected at a single time point. Consequently, we could not control for prior ratings for the models predicting teachers' reports.

Although we aimed to assess EF skills in a more naturalistic *setting*, it is important to note that the EF *tasks* themselves are not naturalistic, at least not in the same way as direct observation of real-life behaviors. In early childhood, some researchers have employed game-like EF tasks that incorporate activities that are common in classroom or playground settings, such as walking on a straight line, using a quiet voice, or waiting for a desirable object. For example, the Head-Toes-Knees-Shoulders (HTKS) task, a highly accessible and fun task with a simple administration procedure, has been widely used to assess and live-code preschoolers' EF skills (Cameron Ponitz et al., 2008). However, these tasks still require a ratio of one child to one assessor and are typically administered in a quiet space outside classrooms (Cameron Ponitz et al., 2008; Carlson, 2005; Carlson & Wang, 2007; Obradović, 2010). Researchers should continue to improve the ecological validity of EF measures and develop more naturalistic, direct measures that better approximate the everyday application of EF skills in educational settings. New tests of EF skills in middle childhood should try to measure the collaborative use of EF skills among peers, using group assessments. While we recognize the limitations of our work and stress the need for continual improvement of ecological validity along these new lines of inquiry, our study nonetheless represents a unique attempt to obtain an objective measure of EF skills in an environment where students are expected to apply EFs regularly.

4.2. Conclusion

By describing and validating a novel method for assessing student EF skills in a classroom setting using tablet computers, we hope to provide researchers with a new way of studying EFs. Relative to the conventional method of a more controlled, one-on-one individual assessment, the group assessment we designed can be administered quickly to many students, requires less assessor training, and measures performance in a more ecologically valid setting. As such, the new protocol represents a pragmatic, cost-effective way to obtain direct assessment of children's EF skills at scale. Although we demonstrate the benefits of testing EFs in a group context, our tablet-based tasks also offer a quick and simple way to assess EFs in an individualized setting. This may be particularly applicable in studies of younger children, who often require more individualized attention and scaffolding during assessment. Our tasks can be easily adjusted for use with different age groups. For example, the program code can be modified to provide younger children more time to respond to each trial, administer additional practice trials, or simplify stimuli (e.g., use only the first five digits in Digit Span Backwards task), if necessary. By using portable computerized tasks that automatically score accuracy and reaction time, researchers can ensure high-quality data collection in large samples by a team of field assessors whose assessment experience may vary.

To facilitate these efforts, we are sharing the tablet apps, source code, and supporting materials with scientists worldwide at no cost using an open-source license. For example, Raver and Morris (2016) employed our tasks in evaluating New York City's Pre-K for All initiative, a full-day, universal pre-kindergarten program that enrolled 68,647 children in the 2015–2016 school year (New York City Department of Education, Division of Early Childhood Education, 2014). They administered the tablet-based assessments in order to efficiently examine the longitudinal EF gains in 1145 preschoolers attending the city-wide program (Morris, 2016; Raver & Morris, 2016). Training teachers to assess student EFs on an ongoing basis creates an opportunity to test whether improvements in EFs are linked to students' academic progress and also to identify how classroom and teaching practices contribute to development of EFs in school-age children. Given recent efforts to incorporate measures of socio-emotional skills in school district accountability systems (Bartolino, Arnold, & LaRocca, 2016) and to build research-practice partnerships between universities and school districts (Wentworth, Carranza, & Stipek, 2016), our protocol offers scalable direct assessment of EFs that can be used to validate and complement student report of self-control (West, 2016).

Acknowledgements

This research was supported by a William T. Grant Foundation Scholar award (180826) to Jelena Obradović. The preparation of this manuscript is also supported by a William R. and Sara Hart Kimball Stanford Graduate Fellowship to Jenna Finch. The authors thank the children, teachers, and school administrators who participated and made this research possible, and many graduate and undergraduate students who helped collect and process the data. The findings, conclusions, and opinions here are those of the authors and do not represent

views of the William T. Grant Foundation, the IES, or the U.S. Department of Education.

References

Allan, N. P., Hume, L. E., Allan, D. M., Farrington, A. L., & Lonigan, C. J. (2014). Relations between inhibitory control and the development of academic skills in preschool and kindergarten: A meta-analysis. *Developmental Psychology*, *50*, 2368–2379. http://dx.doi.org/10.1037/a0037493.

Allison, P. (2009). *Fixed effects regression models*. Thousand Oaks, CA: Sage.

Bartolino, K., Arnold, R., & LaRocca, R. (2016). Expanding the definition of student success: A case study of the CORE districts. Retrieved from http://www.transformingeducation.org/core-toolkit/

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*, 1641–1660. http://dx.doi.org/10.1111/j.1467-8624.2010.01499.x.

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., Ursache, A., Greenberg, M., Vernon-Feagans, L., & Family Life Project Investigators (2015). Multiple aspects of self-regulation uniquely predict mathematics but not letter–word knowledge in the early elementary grades. *Developmental Psychology*, *51*, 459–472. http://dx.doi.org/10.1037/a0038813.

Blankenship, T. L., & Bell, M. A. (2015). Frontotemporal coherence and executive functions contribute to episodic memory during middle childhood. *Developmental Neuropsychology*, *40*, 430–444. http://dx.doi.org/10.1080/87565641.2016.1153099.

Brocki, K. C., & Bohlin, G. (2004). Executive functions in children aged 6 to 13: A dimensional and developmental study. *Developmental Neuropsychology*, *26*, 571–593. http://dx.doi.org/10.1207/s15326942dn2602_3.

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, *8*, 36–41. http://dx.doi.org/10.1111/cdep.12059.

Bush, G., & Shin, L. M. (2006). The Multi-Source Interference Task: An fMRI task that reliably activates the cingulo-frontal-parietal cognitive/attention network. *Nature Protocols*, *1*, 308–313. http://dx.doi.org/10.1038/nprot.2006.48.

California Department of Education (2016, October 4a). California STAR program - 2013 STAR test results (CA Dept of Education). Retrieved from http://star.cde.ca.gov/star2013/index.aspx

California Department of Education (2016, October 4b). Smarter balanced assessment system - testing (CA Dept of Education). Retrieved from http://www.cde.ca.gov/ta/tg/sa/

Cameron Ponitz, C. E., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Child Research Quarterly*, *23*, 141–158. http://dx.doi.org/10.1016/j.ecresq.2007.01.004.

Campbell, S. B., Denham, S. A., Howarth, G. Z., Jones, S. M., Whittaker, J. V., Williford, A. P., ... Darling-Churchill, K. (2016). Commentary on the review of measures of early childhood social and emotional development: Conceptualization, critique, and recommendations. *Journal of Applied Developmental Psychology*, *45*, 19–41. http://dx.doi.org/10.1016/j.appdev.2016.01.008.

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*, *28*, 595–616. http://dx.doi.org/10.1207/s15326942dn2802_3.

Carlson, S. M., & Wang, T. S. (2007). Inhibitory control and emotion regulation in preschool children. *Cognitive Development*, *22*, 489–510. http://dx.doi.org/10.1016/j.cogdev.2007.08.002.

Ciairano, S., Visu-Petra, L., & Settanni, M. (2007). Executive inhibitory control and cooperative behavior during early school years: A follow-up study. *Journal of Abnormal Child Psychology*, *35*, 335–345. http://dx.doi.org/10.1007/s10802-006-9094-z.

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*(11), 2037–2078. http://dx.doi.org/10.1016/j.neuropsychologia.2006.02.006.

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

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237–251. http://dx.doi.org/10.3102/0013189X15584327.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*, 1428–1446. http://dx.doi.org/10.1037/0012-1649.43.6.1428.

Eisenberg, N., Duckworth, A. L., Spinrad, T. L., & Valiente, C. (2014). Conscientiousness: Origins in childhood? *Developmental Psychology*, *50*, 1331–1349. http://dx.doi.org/10.1037/a0030977.

Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC-IV assessment* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Friso-van den Bos, I., van der Ven, S. H., Kroesbergen, E. H., & van Luit, J. E. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, *10*, 29–44. http://dx.doi.org/10.1016/j.edurev.2013.05.003.

Fuhs, M. W., Farran, D. C., & Nesbitt, K. T. (2013). Preschool classroom processes as predictors of children's cognitive self-regulation skills development. *School Psychology Quarterly*, *28*(4), 347–359. http://dx.doi.org/10.1037/spq0000031.

Fuhs, M. W., Farran, D. C., & Nesbitt, K. T. (2015). Prekindergarten children's executive functioning skills and achievement gains: The utility of direct assessments and teacher ratings. *Journal of Educational Psychology*, *107*, 207–221. http://dx.doi.org/10.1037/a0037366.

Hightower, A. D. (1986). The Teacher–Child Rating Scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review*, *15*, 393–409.

Hughes, C. (2011). Changes and challenges in 20 years of research into the development of executive functions. *Infant and Child Development*, *20*, 251–271. http://dx.doi.org/10.1002/icd.736.

Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research*, *85*, 512–552. http://dx.doi.org/10.3102/0034654314561338.

Kochanska, G., & Knaack, A. (2003). Effortful control as a personality characteristic of young children: Antecedents, correlates, and consequences. *Journal of Personality*, *71*, 1087–1112. http://dx.doi.org/10.1111/1467-6494.7106008.

Lagattuta, K. H., Sayfan, L., & Monsour, M. (2011). A new measure for assessing executive function across a wide age range: Children and adults find happy-sad more difficult than day-night. *Developmental Science*, *14*, 481–489. http://dx.doi.org/10.1111/j.1467-7687.2010.00994.x.

Liu, Y., Angstadt, M., Taylor, S. F., & Fitzgerald, K. D. (2016). The typical development of posterior medial frontal cortex function and connectivity during task control demands in youth 8–19 years old. *NeuroImage*, *137*, 97–106. http://dx.doi.org/10.1016/j.neuroimage.2016.05.019.

McCabe, L. A., Hernandez, M., Lara, S. L., & Brooks-Gunn, J. (2000). Assessing preschoolers' self-regulation in homes and classrooms: Lessons from the field. *Behavioral Disorders*, *26*, 53–69.

McCabe, L. A., Rebello-Britto, P., Hernandez, M., & Brooks-Gunn, J. (2004). Games children play: Observing young children's self-regulation across laboratory, home, and school settings. In R. DelCarmen-Wiggins, & A. Carter (Eds.), *Handbook of infant, toddler, and preschool mental health assessment* (pp. 491–521). New York, NY: Oxford University Press.

McClelland, M. M., Acock, A. C., & Morrison, F. J. (2006). The impact of kindergarten learning-related skills on academic trajectories at the end of elementary school. *Early Child Research Quarterly*, *21*, 471–490. http://dx.doi.org/10.1016/j.ecresq.2006.09.003.

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology*, *46*, 235–261. http://dx.doi.org/10.1016/j.jsp.2007.05.001.

McNeish, D., & Wentzel, K. R. (2016). Accommodating small sample sizes in three-level models when the third level is incidental. *Multivariate Behavioral Research*, 1–16. http://dx.doi.org/10.1080/00273171.2016.1262236.

Morris, P. (2016, May). Making universal pre-k work: A partnership approach to quality at scale. *Presented at the Education and Inequality in 21st Century America Conference*. Stanford: CA.

Neuenschwander, R., Röthlisberger, M., Cimeli, P., & Roebers, C. M. (2012). How do different aspects of self-regulation predict successful adaptation to school? *Journal of Experimental Child Psychology*, *113*(3), 353–371. http://dx.doi.org/10.1016/j.jecp.2012.07.004.

New York City Department of Education, Division of Early Childhood Education (2014). Welcome to pre-K. Retrieved from http://schools.nyc.gov/NR/rdonlyres/596B3DEE-2EB5-4C15-A22E-13E952A52E72/0/Prekforall_welcomefull.pdf

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*, 250–256. http://dx.doi.org/10.1037/0022-3514.35.4.250.

Oberle, E., & Schonert-Reichl, K. A. (2013). Relations among peer acceptance, inhibitory control, and math achievement in early adolescence. *Journal of Applied Developmental Psychology*, *34*(1), 45–51. http://dx.doi.org/10.1016/j.appdev.2012.09.003.

Obradović, J. (2010). Effortful control and adaptive functioning of homeless children: Variable-focused and person-focused analyses. *Journal of Applied Developmental Psychology*, *31*, 109–117. http://dx.doi.org/10.1016/j.appdev.2009.09.004.

Obradović, J., Portilla, X. A., & Boyce, W. T. (2012). Executive functioning and developmental neuroscience: Current progress and implications for early childhood education. In R. C. Pianta, L. Justice, S. Barnett, & S. Sheridan (Eds.), *The handbook of early education* (pp. 324–351). New York, NY: Guilford.

Prager, E. O., Sera, M. D., & Carlson, S. M. (2016). Executive function and magnitude skills in preschool children. *Journal of Experimental Child Psychology*, *147*, 126–139. http://dx.doi.org/10.1016/j.jecp.2016.01.002.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development*, *82*, 362–378. http://dx.doi.org/10.1111/j.1467-8624.2010.01561.x.

Raver, C. C., & Morris, P. (2016). Pre-K for all: Snapshot of student learning, executive functioning results. New York, NY. Retrieved from http://www.nyc.gov/html/ceo/downloads/pdf/Westat_Metis_Branch_PreK_Study_Snapshot_of_Student_Learning_Final.pdf

Raykov, T., Dimitrov, D. M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*, 265–279. http://dx.doi.org/10.1080/10705511003659417.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal*, *48*, 335–360. http://dx.doi.org/10.3102/0002831210374874.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology*, *45*, 958–972. http://dx.doi.org/10.1037/a0015861.

Roy, A. L., McCoy, D. C., & Raver, C. C. (2014). Supplemental material for instability versus quality: Residential mobility, neighborhood poverty, and children's self-regulation. *Developmental Psychology*, *50*, 1891–1896. http://dx.doi.org/10.1037/a0036984.supp.

Sasser, T. R., Bierman, K. L., & Heinrichs, B. (2015). Executive functioning and school adjustment: The mediational role of pre-kindergarten learning-related behaviors. *Early Child Research Quarterly*, *30*, 70–79. http://dx.doi.org/10.1016/j.ecresq.2014.09.001.

Sbordone, R. J. (2001). Limitations of neuropsychological testing to predict the cognitive and behavioral functioning of persons with brain injury in real-world settings. *NeuroRehabilitation*, *16*, 199–201.

Schmitt, S. A., McClelland, M. M., Tominey, S. L., & Acock, A. C. (2015). Strengthening school readiness for Head Start children: Evaluation of a self-regulation intervention. *Early Child Research Quarterly*, *30*, 20–31. http://dx.doi.org/10.1016/j.ecresq.2014.08.001.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Silver, C. H. (2014). Sources of data about children's executive functioning: Review and commentary. *Child Neuropsychology*, *20*, 1–13. http://dx.doi.org/10.1080/09297049.2012.727793.

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *The Quarterly Journal of Experimental Psychology*, *59*, 745–759. http://dx.doi.org/10.1080/17470210500162854.

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. http://dx.doi.org/10.1111/j.1750-8606.2011.00209.x.

Ursache, A., Noble, K. G., & Blair, C. (2015). Socioeconomic status, subjective social status, and perceived stress: Associations with stress physiology and executive functioning. *Behavioral Medicine*, *41*, 145–154. http://dx.doi.org/10.1080/08964289.2015.1024604.

Vernon-Feagans, L., Cox, M., & The Family Life Project Key Investigators (2013). The Family Life Project: An epidemiological and developmental study of young children living in poor rural communities. *Monographs of the Society for Research in Child Development*, *78*(5), vii–150.

Weiland, C., Barata, M. C., & Yoshikawa, H. (2014). The co-occurring development of executive function skills and receptive vocabulary in preschool-aged children: A look at the direction of the developmental pathways. *Infant and Child Development*, *23*, 4–21. http://dx.doi.org/10.1002/icd.1829.

Wentworth, L., Carranza, R., & Stipek, D. (2016). A university and district partnership closes the research-to-classroom gap. *Phi Delta Kappan*, *97*(8), 66–69. http://dx.doi.org/10.1177/0031721716647024.

West, M. R. (2016). *Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts.* (Evidence Speaks Reports Vol. 1, No. 13).

Yeniad, N., Malda, M., Mesman, J., van IJzendoorn, M. H., Emmen, R. A., & Prevoo, M. J. (2014). Cognitive flexibility children across the transition to school: A longitudinal study. *Cognitive Development*, *31*, 35–47. http://dx.doi.org/10.1016/j.cogdev.2014.02.004.

Yeniad, N., Malda, M., Mesman, J., van IJzendoorn, M. H., & Pieper, S. (2013). Shifting ability predicts math and reading performance in children: A meta-analytical study. *Learning and Individual Differences*, *23*, 1–9. http://dx.doi.org/10.1016/j.lindif.2012.10.004.