Journal of Cognitive Enhancement
https://doi.org/10.1007/s41465-018-0095-y

ORIGINAL RESEARCH

# Measuring and Improving Executive Functioning in the Classroom

Brian C. Kavanaugh[1] · Omer Faruk Tuncer[2] · Bruce E. Wexler[3]

Received: 4 June 2018 / Accepted: 10 September 2018


## Abstract

Executive function (EF) is a collection of self-regulatory control processes that are compromised by poverty and powerfully predict academic outcomes in children. Despite this, there are few evidence-based interventions to improve EF. Given the importance of measurement of EF in the context of the classroom where children learn, we first report results showing the validity and reliability of over 60,000 web-based, classroom administrations of tests of EF that have previously only been widely used in laboratory research. Using these tests, we next show that 800 min of computer-presented cognitive training exercises can improve EF, after controlling for practice effects and developmental effects (working memory: partial $\eta^2 = .039$, response inhibition: partial $\eta^2 = .132$, interference control: partial $\eta^2 = .072$). The abilities to measure and improve EF at low cost and large scale in classrooms can contribute to improved, evidence-based education and potentially help reduce achievement gaps associated with poverty.

**Keywords** Executive functioning · Cognitive training · Classroom

Executive functioning (EF) is a collection of self-regulatory control processes that are divided into core domains of working memory (i.e., maintain/manipulate data not perceptually present), inhibition (i.e., inhibit or control of attention, thoughts, behaviors), and flexibility (i.e., shift flexibly between tasks/sets; Diamond 2013; Miyake et al. 2000). Other models of attention and/or EF (e.g., Mirsky 1996; Cohen et al. 1998) describe a similar set of cognitive functions albeit in somewhat varying descriptors/classifications (e.g., sustained attention versus inhibitory control, flexibility versus shifting, attention versus interference control). While such functions can be considered 'attention' or 'executive functioning', the term executive functioning (EF) is utilized in the current manuscript.

EF is neurally subserved by the cognitive control network, an interconnected network of frontal, parietal, and subcortical region structures (Senkowski and Gallinat 2015; Niendam et al. 2012). EF is the most vulnerable or sensitive cognitive function to disruption (Diamond 2013), and therefore deficits occur in various childhood clinical conditions (e.g., depression, epilepsy, ADHD) and adverse psychosocial contexts/experiences (e.g., poverty; Evans et al. 2009; Raver et al. 2013).While the most vulnerable, EF is also one of the strongest cognitive predictors of clinical, functional, and academic outcomes (Lee et al. 2013; Baum et al. 2010; Rinsky and Hinshaw 2011; Gligorovic and Durovic 2014).

Particularly relevant to successful childhood outcomes is a child's ability to succeed in the academic environment and EF has been closely tied to childhood academic functioning. The association between EF and academic outcomes has been identified from preschool (Willoughby et al. 2012, 2016) to college/university (Georgiou and Das 2016; Sheehan and Iarocci 2015) in typically developing (Berninger et al. 2017; Best et al. 2011; Cantin et al. 2016; Georgiou and Das 2016; Jacobson et al. 2011; Jacobson et al. 2017) and clinical samples (Biederman et al. 2004; Langberg et al. 2013; Rose et al. 2011; Sirois et al. 2016; Will et al. 2017). EF has been associated with core academic achievement in reading Berninger et al. 2017; Best et al. 2011; Cantin et al. 2016; Georgiou and Das 2016; Jacobson et al. 2017; Rose et al. 2011; Sirois et al. 2016; Will et al. 2017) mathematics (Cantin et al. 2016; Rose et al. 2011; Sirois et al. 2016; Will et al. 2017), science (Latzman et al. 2010), and social studies (Latzman et al. 2010). The EF-academic association extends beyond core

✉ Brian C. Kavanaugh
Brian_Kavanaugh@Brown.edu

[1] E. P. Bradley Hospital/Alpert Medical School of Brown University, East Providence, RI, USA

[2] Ardahan State Hospital, Ardahan, Turkey

[3] Yale University School of Medicine, New Haven, CT, USA

Published online: 28 September 2018

academic development into academic grades (Langberg et al. 2013), history of grade retention (Biederman et al. 2004), academic adjustment/competence (Jacobson et al. 2011; Sheehan and Iarocci 2015), homework problems (Langberg et al. 2013), and academic readiness (Willoughby et al. 2017). Although many studies indicate correlation as opposed to causation, it is without question that EF has a near global association to academic performance throughout childhood.

Despite the critical importance of EF in childhood outcomes, there are few EF-targeted, evidence-based interventions (Diamond and Ling 2016). One area of recent investigation is computerized cognitive training (CCT), which targets cognition and the underlying neural networks to improve clinical and functional outcomes (Keshavan et al. 2014; Diamond and Ling 2016). CCT exerts its effect on cognition through activity-dependent enhancement of neurocognitive systems engaged by increasingly complex cognitive demands (e.g., training the individual to remember more information or with increased distractions). CCT has shown efficacy for many cognitive domains, including learning/memory, processing speed, attention, and EF, across the lifespan in various clinical and healthy populations (Keshavan et al. 2014; Diamond and Ling 2016). A recent meta-analysis (Cortese et al. 2015) found that CCT in children with ADHD resulted in medium to large effect size improvements in working memory and rater-based EF, although no consistent effects were seen in other functions. More recent studies provide additional support that CCT can produce significant improvement in multiple aspects of EF in children, including inhibition, working memory, flexibility, and planning/problem solving (Hadwin and Richards 2016; Van der Donk et al. 2016; Mishra et al. 2016; Zhao et al. 2018). Significant questions have been raised, however, about whether there are far transfer effects of CCT to real-world functional outcomes or even near transfer to laboratory tests dissimilar from the training itself (Cortese et al. 2015). Transfer effects probably depend on features of the specific CCT programs, including, for example, whether they address multiple or single dimensions of EF (Cortese et al. 2015; Diamond and Lee 2011). Initial pilot studies of the multidimensional EF training used in the present study provided evidence of near-transfer improvement in test of EF different from the training exercises and far-transfer to dose-related improvements in reading achievement (Wexler 2013). A more recent and large-sample study showed robust far-transfer effects in improved performance on school-administered math and reading achievement tests (Wexler et al. 2016). While not all training programs have similar functionalities, select programs may have the ability to impact generalized academic outcomes.

In addition, while the potential for CCT to be used efficiently and effectively at large scale has been pointed out (Keshavan et al. 2014; Etkin et al. 2013), this potential has not been developed and evaluated. CCT programs such as CogMed have implemented school-based interventions, although typically in small-group administration outside the classroom and without strong clinical outcomes (Roberts et al. 2016). The validity of EF evaluation within the preschool environment has been established by prior researchers, although evaluations were still conducted in a one-on-one format in a quiet area of the school (Lipsey et al. 2017). Other research is currently examining the efficacy of school-based interventions to improve EF in preschool children (Mind in the Making, Vroom, Circle Time Games; Galinsky et al. 2017). Although evaluation and intervention for EF is emerging in preschool students, there remains a dearth of literature on scalable applications to school-aged children.

The present study addressed both issues of scalability and near transfer, drawing on a large database of pre- and post-intervention EF assessments automatically collected by a commercially available, web-based EF training program used in schools across the USA. Given the value of assessment and intervention within the child's classroom learning environment rather than in a quiet office setting alone with an adult, and the huge national infrastructure provided by schools, this database provides an important research opportunity. Analyses addressed three questions: (1) do formal tests of EF typically used in laboratory research meet embedded checks of performance validity when automatically administered to large numbers of children in classroom settings?; (2) how large are practice effects and test-retest correlations when the tests are administered twice to the same children?; and (3) does a web-based classroom administered EF cognitive skills training program produce improvement in "near-transfer" tests of EF beyond what would be expected from normal developmental gains? Preliminary analyses also examined effects of training time on outcome.

## Method

### Sample/Procedure

All students from kindergarten to grade 8 in schools across the USA who used the described intervention program during the 2014–2015 and 2015–2016 academic years constituted the overall study sample. No exclusion criteria were imposed. The sample of schools is highly diverse in terms of socioeconomic status and types of children within the programs (e.g., special education versus typically developing classrooms). On average in participating schools, 58% of children were receiving free or reduced lunch indicating that the sample contains many children from impoverished backgrounds. As described in the following sections, three distinct hypotheses were tested with the training program database.

As part of the school-based intervention, participants completed baseline evaluations of EF followed by training in 20–

30 min sessions 1–4 times per week for 2–6 months. The intervention is administered in a group format within the classroom setting, sometimes for an entire general class and other times to students selected based on school-assessed special need for cognitive skill training. EF was automatically re-evaluated for all users after approximately 400 and 800 min of CCT, with times varying as a function of student attendance and the proportion of training sessions completed. For a limited period of time, all students were given the tests twice at the beginning of training (i.e., within the first 100 min of training) in order to evaluate test-retest reliabilities and practice effects. These "double-baselines" were not made part of ongoing standard practice in order to lighten demands made of teachers.

The schools purchased the program from the Yale University startup company C8 Sciences which provided training and customer support. Assessments of EF were built into the program by C8 Sciences for purposes of product evaluation and improvement, and the results are provided to the schools to aide in personalizing education. All activities are conducted on school computers within the classroom. All procedures were approved by the Yale University School of Medicine Human Investigations Committee (consent/assent were not required per committee decision since the schools made the decision to use the programs and assessments as part of the school curriculum).

## Assessment and Intervention

**Assessment** Neurocognitive outcomes were assessed with three web-based measures of EF automatically presented, administered, and scored in the classroom setting of the EF training program itself. Two tests followed the design of tests in the NIH Toolbox of tests of executive function (nihtoolbox.org). The first was the Flanker Test. In this task, children have to indicate by keyboard response the pointing direction (right or left) of the center arrow in a linear horizontal array of five arrows. On incongruent trials, the four "flanking" arrows point in the opposite direction of the central arrow. There are 29 congruent and 17 incongruent trials presented in pseudorandom order (with 1–4 congruent trials preceding each incongruent trial and one place where there are 2 incongruent in succession). The flanking arrows precede the central arrow by 100 m, and successive trials are triggered by subject response. The second test was the List Sorting Working Memory Test. Subjects are shown a series of animals or household objects. They then have to click on the objects they have just seen in a grid of 16 objects in order from smallest to largest rather than the order in which they were presented. The test starts with a list of two objects. If the subject completes the list accurately, list length is increased by one. If they err, the same length list is repeated. Two failed attempts at the same list length end the test. The score is the sum of correct list lengths. In part one, trials of animals and household objects alternate. In part two, animals and household objects are presented in the same trial, and subjects have to reorder the animals first and then the household objects. The third test is a go/no-go test of response inhibition. Subjects are instructed to press the space bar whenever a "go" stimulus is presented but not when a "no-go" stimulus is presented. There are three blocks with different stimuli, 50 stimuli per block with 40 go and 10 no-go trials, randomized in sets of 10 with 8 go and 2 no-go in each set. In the first block "P" is the go stimulus and "R" is the no-go stimulus. In the second block this is reversed. In the third block, pictures of furniture are go trials and pictures of foods like cake and ice cream are no-go stimuli. Stimuli are presented for 400 ms with a 1400 ms response window after stimulus offset. Errors are indicated by display of a large red "X."

Consistent with prior research, primary outcome scores were the List Sorting Working Memory Test: total score, go/no-go (GNG) test: nogo condition correctly skipped raw score, and Flanker test: incongruent condition correct response reaction time. Data cleaning involved the identification of those tests that were deemed valid/invalid based on performance validity criteria described below.

**Computer-Presented EF Training Exercises** The brain training games were designed by BEW and developed and supported as web-based applications by the Yale startup company C8Sciences. There are four games each with 80–150 levels of difficulty and two simple spatial span exercises. The first two games have identical underlying computer code and sequence of cognitive challenges but with different user interface game features. These two games were designed to train multiple components of executive functioning; focused attention, response inhibition, cognitive flexibility, divided attention, and working memory. In one, a "magic orb" floats randomly through an underground cave and intermittently turns into a gem. The game begins with the child having to click on red gems, exercising sustained attention. The orb moves faster following correct responses and slows after errors. As they either reach a preset high level of performance or stay at a lower performance level without improvement for an extended period of time, the child is moved through progressive levels that layer in additional cognitive demands. On the next level the orb sometimes turns blue (a foil) that is to be ignored, adding response inhibition. Next, the target color randomly changes back and forth between blue and red gems, increasing required response inhibition and adding cognitive flexibility. Next levels require working memory as half-gems are presented. When two of the same color appear consecutively, the child has to click on the second to complete a whole one. Next, they have to click on a half gem if it is a different color from the one before to create mixed color gems. All rules are repeated with two and then three balls on the screen. The other

game of this pair follows the same sequence of cognitive demands but the child clicks on different types of monkeys, or pirate clothing, as a magic lens jumps around to reveal what is inside moving rows of crates. In the third game, children click on objects that a pirate throws out of a bottomless crate only if the object is a member of a designated category (e.g., animals, furniture, tools, machines) of things the pirate is then looking for. With correct responses, the objects move faster and more objects are on the screen at the same time (from 1 to 6). At higher levels, categories rotate, two categories are targets simultaneously, or the child must find two objects on the screen in the same category. Use of categories is a higher order EF requiring top-down control and organization of information. The game also demands attention, response inhibition, working memory, and cognitive flexibility. The fourth game requires the child to figure out the rule that links a series of three objects and use this rule to choose a fourth object to complete the row. Time to respond becomes shorter with correct responses and rules become more complex in higher levels. This task is designed to train pattern recognition and inductive thinking, as well as attention, response inhibition, and cognitive flexibility. The final two games are a pair of simple spatial memory span games that change only in list length and differ from one another in the visual space in which locations are to be recalled. These tasks are designed to train spatial working memory.

## Statistical Analyses

As described above, the present study addressed questions about test performance validity, practice effects, and test-retest reliability, and intervention effectiveness. Effect sizes were calculated using partial $\eta^2$: small = .01; medium = .06; large = .14. Significance was set at $p < .01$.

**Test Performance Validity** All tests administered during the two school years were evaluated to determine the percentage of test administrations that met embedded performance validity and effort criteria. Literature search revealed that test performance validity are very rarely applied in research studies and that there are no agreed upon criteria. Rationale for the criteria adopted are provided below. Data were available on 19,413 WM, 20,775 GNG, and 25,689 Flanker tests, with variation due to student attendance and test completion.

- WMT validity criteria: WMT total score $\geq 2$ demonstrated that the child understood that they needed to recall the animals or household items they saw and reorder them from smallest to largest when responding.
- GNG validity criteria: (1) Go condition accuracy > 84%, (2), $\leq 10$ trials with response time > 2001 ms, (3) $\leq 15$ trials with response time < 150 ms. Adequate Go condition accuracy is required both to establish the Go response tendency and so as not to artificially inflate no-go correct counts based on simply not paying attention. Response times > 2001 ms are impossible to interpret since new stimuli appear every 1800 ms and responses in less than 150 ms suggest random tapping responses.
- Flanker validity criteria: (1) congruent condition accuracy > 74%, (2) $\leq 4$ incongruent trials with a response time > 4500 ms, (3) $\leq 7$ congruent trials with a response time > 3500 ms, (4) < 4 trials with a response time < 150 ms, (5) $\geq 8$ correct incongruent trials. Individual trials that exceeded the above response time cut-offs were eliminated from the set of test trials before scoring. With chance performance 50% correct, we required 1.5 × chance on the easy and frequent congruent trials to ensure the child understood and attended to the test. Exclusion thresholds for slow trials are about 3 SDs from the overall means and if common suggest lack of engagement with the test.

**Practice Effects** Practice effects and test-retest reliability were evaluated in children that completed valid tests twice within the first 100 min of training and with less than 25 min of training between administrations (mean calendar days between sessions: 24 days): $n = 547$ (GNG), 877 (WMT) and 1056 (Flanker). Repeated measures ANOVA examined differences between testing sessions while Pearson correlation coefficients examined test-retest reliability.

**Intervention Effects** Effects were evaluated in children who had valid tests at three time points: < 100 min of training (baseline), after 300–600 min of training (midpoint), and > 800 min of training. C.1: repeated measures ANOVA with post hoc pairwise comparisons examined changes in each EF test across the three time points to evaluate training outcomes and dose effects. C.2: change in EF as a function of simply getting older was estimated for children in each grade level based on the difference between the average score at the beginning of the school year for children in that grade and the next grade, multiplied by the percent of the year spent in cognitive training. Sample size across grades for GNG ($n$ = 377): kindergarten = 10; 1 = 14; 2 = 65; 3 = 43; 4 = 31; 5 = 24; 6 = 61; 7 = 76; 8 = 53. Sample size across grades for WMT ($n = 467$): kindergarten = 27; 1 = 23; 2 = 90; 3 = 49; 4 = 35; 5 = 38; 6 = 78; 7 = 72; 8 = 55. Sample size across grades for Flanker ($n = 526$): kindergarten = 17; 1 = 25; 2 = 90; 3 = 61; 4 = 36; 5 = 29; 6 = 82; 7 = 102; 8 = 84. Paired sample $t$ tests examined differences between change associated with training and change expected from simply getting older. C.3: follow-up mixed ANOVA were conducted to examine training-related changes in EF in early elementary (K–2), late elementary (3–5), and middle school (6–8) children. The group × time interaction was the primary outcome variable and follow-up simple effects analyses examined differences in

change between groups. Grade was examined given the a priori hypothesis that cohorts may potentially respond differently to interventions. As there were no a priori hypotheses regarding sex, no sex analyses were conducted.

## Results

### Performance Validity Data

Ninety one percent (91.3%) of the 19,413 WMTs met validity criteria. Nearly 73 % (72.8%) of the 20,775 GNG tests met validity criteria. Eighty six percent (86.5%) of the 25,689 Flanker tests met validity criteria. Only valid testing results were included in subsequent analyses.

### Practice Effects

Repeated measures ANOVA found no statistically significant differences between baseline and repeat testing for WMT total score ($F[1, 876] = 4.243$, $p = .040$, partial $\eta^2 = .005$), GNG no go skipped ($F[1, 546] = .490$, $p = .484$, partial $\eta^2 = .001$). A significant difference was detected in Flanker incongruent correct RT ($F[1, 1055] = 9.947$, $p = .002$, partial $\eta^2 = .009$). Descriptive data is provided in Table 1. Test-retest correlations were robust and statistically significant for Flanker ($r = .673$) and no go skipped ($r = .715$), and significant but lower for WMT ($r = .494$).

### Differences Between Baseline, 400 min, and Post-800 min

**Working Memory Test** Four hundred sixty children (grades K-8) had valid tests at all three time points. Sixty six percent ($n = 303$) of the sample was male and the mean grade for the sample = 4.45. WMT total score increased significantly over time (repeated measures ANOVA, $F[2458] = 9.191$, $p < .001$, partial $\eta^2 = .039$). Pairwise comparisons noted statistically significant improvement from baseline to midpoint (−1.67 [95% CI, −2.63 to −.70], $p = .001$) and baseline to post-800 min (−2.08 [95% CI, −3.09 to −1.06, $p < .001$). No improvement was noted from midpoint to post-800 min (−.41 [95% CI, −1.38 to .55, $p = .402$). Results are provided in Table 2.

**Go/No Go Test** Three hundred sixty six children (grades K-8) had valid tests at baseline, at midpoint (between 300 and 600 min), and after at least 800 min of training. Sixty five percent ($n = 239$) of the sample was male and the mean grade for the sample = 4.90 (range = K to 8). No go skipped increased significantly over time (repeated measures ANOVA, $F[2364] = 29.901$, $p < .001$, partial $\eta^2 = .141$). Pairwise comparisons noted statistically significant improvement from baseline to midpoint (−1.48 [95% CI, −2.09 to −.86], $p < .001$), baseline to post-800 min (−2.45 [95% CI, −3.07 to −1.83, $p < .001$), and midpoint to post-800 min (−.97 [95% CI, −1.53 to −.42, $p = .001$). Results are provided in Table 2.

**Flanker Test** Five hundred seven children (grades K-8) had valid tests at baseline, midpoint (between 300 and 600 min), and after at least 800 min of training. Sixty four percent ($n = 322$) of the sample was male and the mean grade for the sample = 4.88 (range = K to 8). Correct incongruent RT decreased significantly over time (repeated measures ANOVA, $F[2505] = 23.412$, $p < .001$, partial $\eta^2 = .085$). Pairwise comparisons noted statistically significant improvement from baseline to midpoint (76.87 [95% CI, 48.30 to 105.44], $p < .001$), baseline to post-800 min (115.68 [95% CI, 82.33 to 149.02, $p < .001$), and midpoint to post-800 min (38.81 [95% CI, 14.60 to 63.01, $p = .002$). Results are provided in Table 2.

**Table 1** Test-retest data

| | *n* | Baseline | Retest | *p* | Partial η2 | Test-retest *R* |
|---|---|---|---|---|---|---|
| WMT total score | 877 | 15.01 (11.26) | 15.81 (11.49) | .- | 040 | .005 .494 |
| GNG nogo skipped | 547 | 16.19 (6.791) | 16.34 (7.094) | .- | 484 | .001 .715 |
| Flanker incon. correct RT | 1056 | 1121.23 (456.89) | 1086.28 (431.53) | .- | 002 | .009 .673 |

*WMT total score*, Working Memory Test total score; *GNG nogo skipped* = go/nogo test no go condition correctly skipped; *Flanker incon. correct RT* = Flanker test incongruent condition correct response reaction time

**Table 2** Differences between baseline, 400 min, and post-800 min

| | *n* | Baseline | Midpoint | Post 800 min | *p* | Partial η2 | Pairwise |
|---|---|---|---|---|---|---|---|
| WMT: total score | 460 | 14.24 (9.81) | 15.91 (10.41) | 16.32 (10.89) | < .001 | .039 | 1 < 2,3 |
| GNG: nogo skipped | 366 | 16.55 (5.98) | 18.03 (6.03) | 19.00 (6.13) | < .001 | .141 | 1 < 2 < 3 |
| Flanker: incon. correct RT | 507 | 976.31 (438.51) | 899.45 (.345.99) | 860.64 (309.64) | < .001 | .085 | 1 < 2 < 3 |

*WMT total score* = Working Memory Test total score; *GNG no go skipped* = go/nogo test no go condition correctly skipped; *Flanker incon. correct RT* = Flanker test incongruent condition correct response reaction time

### Estimated Grade-Related Change Versus Training-Related Change

Training-related change was significantly greater than estimated grade-related changes on all three EF measures: WMT total score ($t[404] = 3.64$, $p < .001$; actual change 5.3 times greater than estimated grade-related change), GNG nogo skipped ($t[312] = 6.84$, $p < .001$; actual change 12.7 times greater than estimated grade-related change), Flanker incongruent correct RT ($t[422] = -6.45$, $p < .001$; actual change 4.95 times greater than estimated grade-related change). Results are provided in Table 3.

### Grade Effects

Mixed ANOVA revealed a significant time × group interaction for incongruent correct RT ($F[4, 1006] = 12.760$, $p < .001$, partial $\eta^2 = .048$). Simple main effects analysis revealed significant changes in early elementary ($F[2503] = 36.221$, $p < .001$, partial $\eta^2 = .126$) and late elementary ($F[2, 503] = 15.616$, $p < .001$, partial $\eta^2 = .058$), and but not in middle school ($F[2, 503] = .066$, $p = .936$, partial $\eta^2 = .000$). Time × group interactions were not significant for WMT ($F[4, 912] = 1.448$, $p = .216$, partial $\eta^2 = .006$), GNG nogo skipped ($F[4, 724] = 1.408$, $p = .230$, partial $\eta^2 = .008$), or GNG go clicked ($F[4, 724] = 1.216$, $p = .303$, partial $\eta^2 = .007$).

## Discussion

This study demonstrated the feasibility and validity of web-based, computer-presented, and classroom-administered assessment of working memory, response inhibition, and interference control, executive cognitive functions (EF) important in learning which have typically been assessed in more costly office-based clinical and research studies. In addition, using these assessments, we demonstrated that children in K-8 who used a novel computerized cognitive training program showed significant improvement in these key cognitive skills.

### EF Assessment

The Flanker Test and the List Sort Working Memory Test followed descriptions of tests in the NIH-Toolbox of best practice measures of EF. The toolbox does not include a measure of response inhibition, and given the importance of this aspect of EF in predicting school success (Allan et al. 2014; Latzman et al. 2010), a web-based version of the widely used research go/no-go (GNG) test was created. All tests had a "game-like" feel and instructions that include checks and practice to help make sure the child understands how to proceed, with the instructions and practice trials repeated if the first practice trials are incorrect. Still, given the classroom administration of the tests, we thought it was important to apply performance validity checks to eliminate tests where performance may have been compromised due to lack of understanding of the test, lack of effort, or distraction during testing. Considering both accuracy checks and individual trials with unusually short or long response times, 91% of working memory, 86% of Flanker, and 73% of GNG tests were considered valid. There is a dearth of literature utilizing a priori performance validity exclusion criteria for EF tasks in children (and more generally on all neurocognitive measures). Most commonly, researchers have excluded scores greater than 2–3 standard deviations from the test mean. More recent studies have utilized response accuracy and reaction time ranges to remove those scores thought to reflect invalid performance potentially due to factors such as limited comprehension of task expectations and overall effort/motivation (Zabel et al. 2009; Tarp et al. 2016; Zelazo et al. 2013;

**Table 3** Differences between actual and predicted change

| | *n* | Actual change | Predicted change | *p* | Cohen's *D* |
|---|---|---|---|---|---|
| WMT total score | 405 | 2.44 (10.99) | .46 (.66) | < .001 | .18 |
| GNG nogo skipped | 313 | 2.54 (6.05) | .20 (.45) | < .001 | .39 |
| Flanker incongruent correct RT | 423 | − 146.38 (385.78) | − 29.57 (76.55) | < .001 | .31 |

*WMT total score* = Working Memory Test total score; *GNG no go skipped* = go/nogo test no go condition correctly skipped; *Flanker: incon. correct RT* = Flanker test incongruent condition correct response reaction time

Graham et al. 2016). Rates of valid task performance were reported as 82% (Graham et al. 2016) and 92% (Tarp et al. 2016) in two recent child studies. Current rates (i.e., 73%, 86%, 91%) are consistent with these reports of laboratory-based performance validity criteria rates and suggest that valid measurement of EF can be conducted in the classroom environment.

Test-retest reliability with classroom administration on the GNG and Flanker tests fell in the acceptable range ($r$ = .67–.73), while reliability on the working memory test was at the upper end of the poor range ($r$ = .49). Although the NIH Toolbox Flanker and List Sorting Working Memory tasks in children show good to excellent test-retest reliability ($r$ = .86–.92; Tulsky et al. 2013; Zelazo et al. 2013), other test-retest reliability data in children have shown more modest results consistent with the current findings (ranging from .09 to .87 for EF tasks in children; Leark et al. 2004; Soreni et al. 2009; Zabel et al. 2009; Bezdjian et al. 2009; Tarp et al. 2016). Adult EF task data has also shown similarly modest findings (ranging from .11 to .88; Langenecker et al. 2007; Bollini et al. 2000; Hahn et al. 2011; Paap and Sawi 2016). Taken together, current test-retest reliability approximates those seen in prior studies utilizing well-established EF measurements in laboratory settings.

Assessment of EF in the settings of research laboratories or clinician offices with a single child sitting with an adult in a quiet room may yield somewhat higher test-retest consistency, but assessment in the classroom setting where the child has to learn has important ecological validity. Moreover, the web-based classroom testing is easily scaled and available at a small fraction of the cost of office-based testing, making it potentially available to much larger numbers of children. Valid, feasible, and scalable classroom testing of EF, as our data indicates is possible, would be of great value in identifying children with special needs or special talents, and providing teachers with information to better personalize education. Moreover, because of the low cost and easy availability of the tests, invalid tests can be repeated after additional instruction, unexpected results can be confirmed with retesting, and teachers can assess effects of their interventions with pre- and post-intervention assessments. Equally important, valid measurement of EF in large numbers of children in classroom settings makes possible new research on development and significance of EF, and on classroom interventions to improve EF. Such research is essential to the development of evidence-based pedagogy.

## Effectiveness of EF Training

Results of the current study suggest that EF training within the classroom setting can lead to statistically significant and meaningful EF improvement, and contribute to the growing body of research that suggest cognitive training interventions can result in EF improvements in children (Cortese et al. 2015; Diamond and Ling 2016; Hadwin and Richards 2016; Van der Donk et al. 2016; Mishra et al. 2016; Zhao et al. 2018). Collecting naturalistic data from real-world implementations of the EF training program makes possible large study samples without research funds typically needed to gather such data and offers the important advantage of assessment of outcomes in a range of more general use settings than is often the case. However, the absence of a randomized control group raises possibilities of at least three potentially significant confounds. First, results are simple practice effects from taking the EF assessments two times. To address this possible confound, we evaluated improvement in samples of 500 to 1000 children taking the tests twice with little training or time between the two tests. Whatever practice effects are present in these samples are expected to be greater than in our implementation outcome data where the test and retests were separated by much longer time intervals. Practice effects were small and statistically insignificant for working memory and interference control. Practice effects were statistically significant for response inhibition but an order of magnitude smaller than the observed improvement after training (partial $\eta 2$ = .009 versus .141).

The second possible confound is that EF may have improved in the implementation sample simply because the children got older, and during this time had the stimulation offered by their school programs. Fortunately, our sample allowed estimate of the expected time-related improvement in EF. Using EF test scores from the beginning of the school year for each grade, we used the difference from each year to the next as the 12-month time-related improvement from biological maturation and general environmental stimulation. Using the actual calendar time between tests at baseline and after 800-min training to estimate the expected time-related improvement in test scores for that child, we found change observed during the intervention was between 4.95 and 12.7 times greater than predicted neurodevelopmental growth during the same time period ($d$ = .18 to .31).

The third possible confound is that the schools using the intervention also introduced other program changes, and that these program elements rather than the EF cognitive training program were the cause of the observed improvement in EF. While we cannot rule this possibility out, three things make it likely that the improvements were related to the EF training intervention. First is the fact that the EF intervention was designed and employed with the a priori predictions that it would improve EF performance. While it is important to consider alternative explanations of the improvements that were then observed, it is reasonable to think they are likely due to the program that was specifically designed to produce the improvements. Secondly, the improvements were evident in all grade levels from K-8 and across many schools and classrooms across the country. The only thing all these schools and

classrooms had in common was the use of the EF skills training program. It is actually exceedingly unlikely that they all also introduced other programs that were the source of observed improvement in EF. Finally, we found a dose response for improvements in inhibition and interference control, and such dose response relationships are thought to further link the response to the intervention. Of note, while no grade effect was found for working memory or response inhibition, significant gains in interference control were observed in younger, but not older children. Interference control was measured by reaction time on correct incongruent trials, and this indicator continues to improve as children develop through the full age range of our study population. However, several aspects of neurocognitive function contribute to better scores on this indicator, including, for example, focused attention and motor speed. Perhaps Activate training impacts aspects of function that are developing at earlier ages, but not factors that are the primary basis of the improvement at later ages.

Current results only report on the direct effects of the EF tasks, without examination of far transfer or ecologically valid findings (e.g., grades, academic test performance). In a recent related study, Wexler et al. (2016) identified the currently examined cognitive training program (three sessions per week for 4 months) showed significant effects on school-administered math and reading achievement tests. While the current study utilized a 'target engagement' approach (i.e., can the program modulate the primary target of intervention), these recent results provide evidence to suggest that this target engagement can lead to subsequent enhancement of academic outcomes.

Current results have critical implications for child developmental outcomes. EF deficits are the most common type of cognitive deficits and can occur within the context of any type of neurodevelopmental disruption, such as onset of a clinical condition or living in adverse or suboptimal environments (Diamond 2013). Despite the fact that EF is often the strongest predictor of clinical/functional outcomes in these children, there are very few evidence-supported and scalable interventions to improve EF. Not only do current results suggest this cognitive training program can effectively improve EF, but it can validly measure pre-/post-training EF and do so within the child's primary learning environment. These findings have direct implications for socioeconomically disadvantaged children. Given the influence of EF on academic outcomes, cognitive training that successfully targets and improves EF has the potential to help reduce the large achievement gaps associated with poverty (Reardon 2013; Bohrnstedt et al. 2015), and indirectly with race, across the USA (Blair and Raver 2016). School-based EF evaluation and intervention also holds promise as a primary and/or secondary preventative intervention for EF deficits related to clinical neurodevelopmental disorders.

Although demographic data on individual students is limited, the majority (58%) of students in this sample were eligible for the National School Lunch Program, an indicator of poverty. Analysis of the relationship between poverty in a larger sample that includes the present sample found robust relationships between percent of children NSLP eligible and the same tests of inhibition and working memory used in this study (Imal et al., submitted). Repeated unpublished analysis of data from individual schools with more than 50% of the children NSLP eligible have shown both school wide gains in attention, self-control and memory, and greater gains in children who were weak at baseline. On the other hand, our published report (Wexler et al. 2016) of far-transfer training effects to improvement on school-administered math and reading tests was from schools with less than 10% of children NSLP eligible. Unpublished data provided by other schools, however, has shown even greater gains on math and reading proficiency tests in general classroom implementations of the program in poverty community schools. This data together, then, not only shows gains in students from both affluent and impoverished communities, but also suggests the gains are larger in the later part. Regarding children with identified learning disorders, teachers have provided multiple anecdotal reports of substantial improvement, but the absence of more complete demographic and clinical data has prevented systematic analysis.

## Limitations

As noted above, the absence of a non-intervention control group is a significant limitation, and while major potential confounds have been addressed to a significant degree, the evidence does not meet the highest evidentiary standard of a randomized controlled study, and follow-up randomized clinical trials would be of value. Current results should be considered correlational data and not causal findings, given the lack of randomization and possibility that observed improvements were secondary to other program initiatives in the schools using the EF training program. Very limited clinical/demographic information was available on individual children in the current study, and it will be important in future research to better understand user characteristics that predict response to the EF training intervention. Finally, variability in the number of children in each school and grade prevents use of sophisticated statistical models (e.g., nested data), and variability in the types of students across grades/schools further limits interpretability of grade effects.

## Compliance with Ethical Standards

**Conflict of Interest** Bruce E. Wexler reports a financial interest in C8 Sciences, a Yale startup company that sells the Activate program to schools. The other authors declare that they have no conflict of interest.

**Ethical Approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

**Informed Consent** All procedures were approved by the Yale University School of Medicine Human Investigations Committee (consent/assent were not required per committee decision since the schools made the decision to use the programs and assessments as part of the school curriculum).

## 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.

Baum, K. T., Byars, A. W., dGrauw, T. J., Dunn, D. W., Bates, J. E., Howe, S. R., Chiu, C. Y., & Austin, J. K. (2010). The effect of temperament and neuropsychological functioning on behavior problems in children with new-onset seizures. *Epilepsy & Behavior, 17*, 467–473.

Berninger, V., Abbott, R., Cook, C. R., & Nagy, W. (2017). Relationships of attention and executive functions to oral language, reading, and writing skills and systems in middle childhood and early adolescence. *Journal of Learning Disabilities, 50*, 434–449.

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences, 21*, 327–336.

Bezdjian, S., Baker, L. A., Lozano, D. I., & Raine, A. (2009). Assessing inattention and impulsivity in children during the go/nogo task. *British Journal of Developmental Psychology, 27*, 365–383.

Biederman, J., Monuteaux, M. C., Doyle, A. E., Seidman, L. J., Wilens, T. E., Ferrero, F., Morgan, C. L., & Faraone, S. V. (2004). Impact of executive function deficits and attention-deficit/hyperactivity disorder (ADHD) on academic outcomes in children. *Journal of Consulting and Clinical Psychology, 72*, 757–766.

Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., & Chan, D. (2015). School composition and the black-white achievement gap. NCES 2015-018. National Center for education statistics. https://nces.ed.gov/pubsearch. Accessed on April 2018

Bollini, A. M., Arnold, M. C., & Keefe, R. S. (2000). Test-retest reliability of the dot test of visuospatial working memory in patients with schizophrenia and controls. *Schizophrenia Research, 45*, 169–173.

Cantin, R. H., Gnaedinger, E. K., Gallaway, K. C., Hesson-McInnis, M. S., & Hund, A. M. (2016). Executive functioning predicts reading, mathematics, and theory of mind during elementary years. *Journal of Experimental Child Psychology, 146*, 66–78.

Cohen, R. A., Malloy, P. F., & Jenkins, M. A. (1998). Disorders of attention. In P. J. Snyder & P. D. Nussbaum (Eds.), *Clinical neuropsychology: A pocket handbook for assessment*. Washington DC: American Psychological Association.

Cortese, S., Ferrin, M., Brandeis, D., Buitelaar, J., Daley, D., Dittmann, R. W., Holtmann, M., Santosh, P., Stevenson, J., Stringaris, A., Zuddas, A., & Sonuga-Barke, E. J. (2015). Cognitive training for attention-deficit/hyperactivity disorder: Meta-analysis of clinical and neuropsychological outcomes from randomized controlled trials. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*, 164–174.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135–168.

Diamond, A., & Lee, K. (2011). Interventions and programs demonstrated to aid executive function development in children 4-12 years of age. *Science, 333*, 959–964.

Diamond, A., & Ling, D. S. (2016). Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not. *Developmental Cognitive Neuroscience, 18*, 34–48.

Etkin, A., Gyurak, A., & O'Hara, R. (2013). A neurobiological approach to the cognitive deficits of psychiatric disorders. *Dialogues in Clinical Neuroscience, 15*, 419–429.

Evans, G. W., Schamberg, M. A., & McEwen, B. S. (2009). Childhood poverty, chronic stress, and adult working memory. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 6545–6549.

Galinsky, E., Bezos, J., McClelland, M., Carlson, S. M., & Zelazo, P. D. (2017). Civic science for public use: Mind in the making and vroom. *Child Development, 88*, 1409–1418.

Georgiou, G. K., & Das, J. P. (2016). What component of executive functions contributes to normal and impaired reading comprehension in young adults? *Research in Developmental Disabilities, 49-50*, 118–128.

Gligorovic, M., & Durovic, N. B. (2014). Inhibitory control and adaptive behavior in children with mild intellectual disability. *Journal of Intellectual Disability Research, 58*, 233–242.

Blair, C., & Raver, C. C. (2016). Poverty, stress, and brain development: New directions for prevention and intervention. *Academic Pediatrics, 16*, s30–s36.

Graham, D. M., Glass, L., & Mattson, S. N. (2016). The influence of extrinsic reinforcement on children with heavy prenatal alcohol exposure. *Alcoholism: Clinical and Experimental Research, 40*, 348–358.

Hadwin, J. A., & Richards, H. J. (2016). Working memory training and CBT reduces anxiety symptoms and attentional biases to threat: A preliminary study. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00047.

Hahn, E., Ta, T. M., Hahn, C., Kuehl, L. K., Ruehl, C., Neuhaus, A. H., & Dettling, M. (2011). Test-retest reliability of attention network test measures in schizophrenia. *Schizophrenia Research, 133*, 218–222.

Jacobson, L. A., Williford, A. P., & Pianta, R. C. (2011). The role of executive function in children's competent adjustment to middle school. *Child Neuropsychology, 17*, 255–280.

Jacobson, L. A., Koriakin, T., Kipkin, P., Boada, R., Frijters, J. C., Lovett, M. W., Hill, D., Willcutt, E., Gottwald, S., Wolf, M., Bosson-Heenan, J., Gruen, J. R., & Mahone, E. M. (2017). Executive functions contribute uniquely to reading competence in minority youth. *Journal of Learning Disabilities, 4*, 1–12.

Keshavan, M. S., Vinogradov, S., Rumsey, J., Sherrill, J., & Wagner, A. (2014). Cognitive training in mental disorders: Update and future directions. *American Journal of Psychiatry, 171*, 510–522.

Langberg, J. M., Dvorsky, M. R., & Evans, S. W. (2013). What specific facets of executive function are associated with academic functioning in youth with attention-deficit/hyperactivity disorder? *Journal of Abnormal Child Psychology, 41*, 1145–1159.

Langenecker, S. A., Zubieta, J.-K., Young, E. A., Akil, H., & Nielson, K. A. (2007). A task to manipulate attentional load, set-shifting, and inhibitory control: Convergent validity and test-retest reliability of the parametric go/no-go test. *Journal of Clinical and Experimental Neuropsychology, 29*, 842–853.

Leark, R. A., Wallace, D. R., & Fitzgerald, R. (2004). Test-retest reliability and standard error of measurement for the test of variables of attention (TOVA) with healthy school-age children. *Assessment, 11*, 285–289.

Lee, R. S., Hermens, D. F., Redoblado-Hodge, M. A., Naismith, S. L., Porter, M. A., Kaur, M., White, D., Scott, E. M., & Hickie, I. B. (2013). Neuropsychological and socio-occupational functioning in

young psychiatric outpatients: A longitudinal investigation. *PLoS One, 8*, e58176. https://doi.org/10.1371/journal.pone.0058176.

Latzman, R. D., Elkovitch, N., Young, J., & Clark, L. A. (2010). The contribution of executive functioning to academic achievement among male adolescents. *Journal of Clinical and Experimental Neuropsychology, 32*, 455–462.

Lipsey, M. W., Nesbitt, K. T., Farran, D. C., Dong, N., Fuhs, M. W., & Wilson, S. J. (2017). Learning-related cognitive self-regulation measures for prekindergarten children: A comparative evaluation of the educational relevance of selected measures. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000203.

Mirsky, A. (1996). Disorders of attention: A neuropsychological perspective. In L. G. Reid & A. Norman (Eds.), *Attention, memory, and executive function* (pp. 71–95). Baltimore: Paul H Brookes Publishing.

Mishra, J., Sagar, R., Joseph, A. A., Gazzaley, A., & Merzenich, M. M. (2016). Training sensory signal-to-noise resolution in children with ADHD in a global mental health setting. *Translational Psychiatry, 6*, e781.

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.

Niendam, T. A., Laird, A. R., Ray, K. I., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience, 12*, 241–268.

Paap, K. R., & Sawi, O. (2016). The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods, 274*, 81–83.

Raver, C. C., McCoy, D. C., & Lowenstein, A. L. (2013). Predicting individual differences in low-income children's executive control from early to middle childhood. *Developmental Science, 16*, 394–408.

Reardon, S. F. (2013). The widening income achievement gap. *Educational Leadership, 70*, 10–16.

Rinsky, J. R., & Hinshaw, S. P. (2011). Linkages between childhood executive functioning and adolescent social functioning and psychopathology in girls with ADHD. *Child Neuropsychology, 17*, 368–390.

Roberts, G., Quach, J., Spencer-Smith, M., Anderson, P. J., Gathercole, S., Gold, L., Sia, K.-L., Mensah, F., Rickards, F., Ainley, J., & Wake, M. (2016). Academic outcomes 2 years after working memory training for children with low working memory: A randomized clinical trial. *JAMA Pediatrics, 170*, e154568.

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2011). Modeling a cascade of effects: The role of speed and executive functioning in preterm/full-term differences in academic achievement. *Developmental Science, 14*, 1161–1175.

Senkowski, D., & Gallinat, J. (2015). Dsyfunctional prefrontal gamma-band oscillations reflect working memory and other cognitive deficits in schizophrenia. *Biological Psychiatry, 77*, 1010–1019.

Sheehan, W. A., & Iarocci, G. (2015). Executive functioning predicts academic but not social adjustment to university. *Journal of Attention Disorders*, 1–9. https://doi.org/10.1177/1087054715612258

Sirois, P. A., Chernoff, M. C., Malee, K. M., Garvie, P. A., Harris, L. L., Williams, P. L., Woods, S. P., Nozyce, M. L., Kammerer, B. L., Yildirim, C., & Nichols, S. L. (2016). Associations of memory and executive functioning with academic and adaptive functioning among youth with perinatal HIV exposure and/or infection. *Journal of the Pediatric Infectious Diseases Society, 5*, S2–S32.

Soreni, N., Crosbie, J., Ickowicz, A., & Schachar, R. (2009). Stop signal and conners' continuous performance tasks: Test-retest reliability of two inhibition measures in ADHD children. *Journal of Attention Disorders, 13*, 137–143.

Tulsky, D. S., Carlozzi, N. E., Chevalier, N., Espy, K. A., Beaumont, J. L., & Mungas, D. (2013). NIH toolbox cognition battery (cb): Measuring working memory. *Monographs of the Society for Research in Child Development, 78*, 70–87.

Tarp, J., Domazet, S. L., Froberg, K., Hillman, C. H., Andersen, L. B., & Bugge, A. (2016). Effectiveness of a school-based physical activity intervention on cognitive performance in Danish adolescents: Locomotion-learning, cognition, and motion – A cluster randomized controlled trial. *PLoS One, 11*, e0158087. https://doi.org/10.1371/journal.pone.0158087.

Van der Donk, M. L., Hiemstra-Beernink, A., Tjeenk-Kalff, A. C., van der Leij, A., & Lindauer, R. J. (2016). Predictors and moderators of treatment outcome in cognitive training for children with ADHD. *Journal of Attention Disorders*. https://doi.org/10.1177/1087054716632876.

Wexler, B. E. (2013). Integrated brain and body exercises for ADHD and related problems with attention and executive function. *International Journal of Gaming and Computer-Mediated Simulations, 5*, 1–17.

Wexler, B. E., Iseli, M., Leon, S., Zaggle, W., Rush, C., Goodman, A., Esat Imal, A., & Bo, E. (2016). Cognitive priming and cognitive training: Immediate and far transfer to academic skills in children. *Scientific Reports, 6*, 32859. https://doi.org/10.1038/srep32859.

Willoughby, M. T., Blair, C. B., Wirth, R. J., Greenberg, M., & The Family Life Project Investigators. (2012). The measurement of executive function at age5: Psychometric properties and relationship to academic achievement. *Psychological Assessment, 24*, 226–239.

Willoughby, M. T., Magnus, B., Vernon-Feagans, L., Blair, C. B., & the Family Life Project Investigators. (2017). Developmental delays in executive function from 3 to 5 years of age predict kindergarten academic readiness. *Journal of Learning Disabilities, 50*, 359–372.

Will, E., Fidler, D. J., Daunhauer, L., & Gerlach-McDonald, B. (2017). Executive function and academic achievement in primary grade students with down syndrome. *Journal of Intellectual Disability Research, 61*, 181–195.

Zabel, T. A., von Thomsen, C., Cole, C., Martin, R., & Mahone, E. M. (2009). Reliability concerns in the repeated computerized assessment of attention in children. *The Clinical Neuropsychologist, 23*, 1213–1231.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). NIH toolbox cognition battery (cb): Measuring executive function and attention. *Monographs of the Society for Research in Child Development, 78*, 16–33.

Zhao, X., Chen, L., & Maes, J. H. (2018). Training and transfer effects of response inhibition training in children and adults. *Developmental Science*, 1–12. https://doi.org/10.1111/desc.12511