Journal of the International Neuropsychological Society (2020), **26**, 725–732
Copyright © INS. Published by Cambridge University Press, 2020.
doi:10.1017/S135561772000017X

BRIEF COMMUNICATION

Preliminary Psychometrics for the Executive Function Challenge Task: A Novel, "Hot" Flexibility, and Planning Task for Youth

Lauren Kenworthy[1,*], Andrew Freeman[2], Allison Ratto[1], Katerina Dudley[3], Kelly K. Powell[4], Cara E. Pugliese[1], John F. Strang[1], Alyssa Verbalis[1] and Laura G. Anthony[1,5]

[1]Center for Autism Spectrum Disorders – Children's National Medical Center, Rockville, MD, USA
[2]University of Nevada, Las Vegas, NV, USA
[3]Department of Psychology and Neuroscience – University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[4]Yale Child Study Center – Yale School of Medicine, New Haven, CT, USA
[5]Department of Psychiatry – University of Colorado School of Medicine; Pediatric Mental Health Institute – Children's Hospital of Colorado, Aurora, CO, USA

(Received September 10, 2019; Final revision January 14, 2020; Accepted January 21, 2020; First Published Online March 17, 2020)

Abstract

Objective: Executive functions (EF) drive health and educational outcomes and therefore are increasingly common treatment targets. Most treatment trials rely on questionnaires to capture meaningful change because ecologically valid, pediatric performance-based EF tasks are lacking. The Executive Function Challenge Task (EFCT) is a standardized, treatment-sensitive, objective measure which assesses flexibility and planning in the context of provocative social interactions, making it a "hot" EF task. **Method:** We investigate the structure, reliability, and validity of the EFCT in youth with autism (Autism Spectrum Disorder; $n = 129$), or attention deficit hyperactivity disorder with flexibility problems ($n = 93$), and typically developing (TD; $n = 52$) youth. **Results:** The EFCT can be coded reliably, has a two-factor structure (flexibility and planning), and adequate internal consistency and consistency across forms. Unlike a traditional performance-based EF task (verbal fluency), it shows significant correlations with parent-reported EF, indicating ecological validity. EFCT performance distinguishes youth with known EF problems from TD youth and is not significantly related to visual pattern recognition, or social communication/understanding in autistic children. **Conclusions:** The EFCT demonstrates adequate reliability and validity and may provide developmentally appropriate, treatment-sensitive, and ecologically valid assessment of "hot" EF in youth. It can be administered in controlled settings by masked administrators.

Keywords: Autism, Behavior, ADHD, Cognition, Objective, Measurement

INTRODUCTION

Executive functions (EF) govern the capacity to regulate thinking, behavior, and emotions. They are essential to learning, mental and physical health, and other key adult outcomes (e.g., Moffitt et al., 2011; Snyder, Miyake & Hankin, 2015). As such, they are increasingly targeted in intervention and biomarker research, yet meaningful measurement of EF has been challenging. Performance-based measures of EF are criticized for lacking ecological validity (e.g., Burgess et al., 2006) and often fail to capture EF where it is most in demand: unstructured real-world situations where plans must be formed and flexibly implemented to achieve goals (Chevignard et al., 2000; Toplak et al., 2013). Indeed, traditional performance-based tasks have been identified as, by their nature, ill-suited to replicate real-world EF demands (e.g., Holmes-Bernstein & Waber, 1990), because they make expectations explicit (e.g., "work as quickly as you can") instead of implicit and present problems in a highly controlled setting *versus* a realistic social interaction. Informant report measures of EF often yield ecologically valid data but cannot be used for treatment-masked data collection when informants are aware of the treatment being provided. Furthermore, performance tasks may allow a more precise investigation than informant report tools of specific EF domains, such as flexibility (Miyake et al., 2000). As such, performance tasks are an essential component of treatment trials and biomarker studies.

*Correspondence and reprint requests to: Lauren Kenworthy, Ph.D., Center for Autism Spectrum Disorders, Children's National Medical Center, 15245 Shady Grove Road, Suite 350, Rockville, MD 20850, USA. Email: lkenwort@childrensnational.org

https://doi.org/10.1017/S135561772000017X Published online by Cambridge University Press

Measures of EF are also dichotomized as "hot" *versus* "cool." Traditional performance tasks are described as "cool" when they are decontextualized and lack strong affective or motivational components. This contrasts with "hot" EF measures, which are considered more relevant to real-world decision-making because they emphasize social, motivational, and emotional saliency (Zelazo & Cunningham, 2007). It is more difficult to defer gratification, be flexible, and follow plans when there are strong social expectations or feelings involved (Prencipe et al., 2011). EF performance tasks typically assess "cool" cognitive control functions, which are of less relevance to the expression of EF in psychopathology. They are also of less relevance in real-world situations which elicit arousal/emotional valence. Therefore, "hot" EF measures have been developed for adults to introduce emotional saliency to EF tasks through external rewards (e.g., gambling tasks), but there is emerging evidence that these tasks are not associated with real-world outcomes when adapted for children (Poland, Monks, & Tsermentseli, 2014).

Among elementary school-aged youth, social interactions present some of the greatest EF demands (Strang et al., 2017). They require youth to generate plans for activities, manage multiple inputs simultaneously, integrate information about peers and their personalities, regulate their behavior and emotions, and flexibly shift plans when needed. They are inherently less structured, and more open, than the closed system experiences created in EF laboratory tasks which provide limited response options. A socially mediated EF task that measures a youth's ability to plan and flexibly implement plans is promising as an ecologically valid measure of EF for several reasons. Tapping EF in the context of a provocative social interaction imposes real-world EF demands in a controlled setting, while also increasing the motivational and emotional saliency of the task, making it "hot." Such a model is consistent with the well-developed assessment of self-regulatory dimensions of temperament in infants and pre-school age children using the Laboratory Temperament Assessment Battery (Goldsmith and Rothbart, 1996).

Here, we present preliminary data on the Executive Function Challenge Task (EFCT), which measures flexibility and planning in the context of a social interaction. Flexibility has been identified as a core EF component that underlies more complex EF abilities, such as planning (Friedman & Miyake, 2017). Flexibility and planning are commonly observed problems in autism (American Psychiatric Association, 2013) and other neurodevelopmental (Willcutt, Doyle, Nigg, Faraone, & Pennington, 2005) and acquired disorders (Ozga, Povroznik, Engler-Chiurazzi, & Vonder Haar, 2018). Flexibility is also a common problem in mood and anxiety disorders (Snyder, Miyake & Hankin, 2015). In keeping with the theory that EF is both a unified and diverse construct with separable, meaningful subdomains (Miyake et al., 2000), there is evidence of variability in the profiles of EF difficulties across different disorders (Gioia et al, 2002), as well as specificity in the domains of EF which can be targeted by treatments (Kenworthy et al, 2014). For these reasons, the identification of separate flexibility and planning domains in the EFCT is of interest.

The EFCT's use of a standardized, semi-structured protocol which does not provide explicit rules for completing the tasks creates an open system that mimics the implicit, unspoken expectations for EF in everyday life. By combining developmentally appropriate, and socially and emotionally salient, demands with open-ended tasks, the EFCT is designed to mimic EF challenges to regulate behavior, thinking, and emotions that children encounter in their daily lives. The EFCT was developed for use in EF treatment trials for children with Autism Spectrum Disorder (ASD) or attention deficit hyperactivity disorder (ADHD), in which it showed treatment-specific sensitivity to change (Kenworthy et al., 2014). This paper represents an initial assessment of the EFCT's psychometric properties in a sample combined from two previous treatment trials and two phenotyping studies. Overall, we hypothesize that the EFCT will:

1. Separate into flexibility and planning factors.
2. Have equivalency across Forms A and B.
3. Demonstrate internal reliability, and convergent and discriminant validity by:
   a) Discriminating youth with ASD or ADHD from typically developing (TD) youth.
   b) Correlating with parent report and performance task measures of EF (convergent validity), but not with measures of vocabulary, matrix reasoning, or social communication/understanding (discriminant validity).
   c) Showing a distinct pattern of significant relationships (convergent validity) in which EFCT flexibility scores relate to parent-reported flexibility and verbal fluency switching scores; and EFCT planning scores relate to semantic verbal fluency scores and parent-reported planning/organization.
   d) Regarding validity hypotheses: see arrows in Table 2 which specify all of our predicted relationships.

METHODS

Participants

Two hundred seventy-four 7–18 year olds (129 with ASD, 93 with ADHD, and 52 with TD) with a prorated Wechsler Full-Scale IQ score (FSIQ) ≥ 75 participated in one of four IRB approved studies (Children's National Institutional Review Board). Participants with ASD or ADHD met DSM-5 diagnostic criteria as determined by an experienced clinical psychologist and cutoff criteria on the Autism Diagnostic Observation Schedule, Module 3 first or second edition (ADOS) for ASD (for ASD group: total ADOS score mean = 13.5 ± 5.1; social affect = 10.5 ± 4.0; restricted and repetitive behavior = 3.0 ± 2.2), or the Mini International Neuropsychiatric Interview-Kid for ADHD. All ADHD participants also had flexibility problems as reported by parents or teachers. TD youth had no history of DSM diagnoses in themselves or first-degree relatives, and no psychiatric medication use. See Table 1 for participant demographics.

https://doi.org/10.1017/S135561772000017X Published online by Cambridge University Press

Table 1. Demographic and clinical characteristics (mean score (SD), except as noted)

	Overall ($n = 274$)	ASD ($n = 129$)	ADHD ($n = 93$)	TD ($n = 52$)	Group comparisons	η^2
Age[abc] (years)	10.3 (1.9)	10.3 (1.8)	9.6 (0.9)	11.7 (2.8)	$F(2, 14614.69)$[d] $= 58.90$, $p < .001$	.31
Female, *n* (%)	61 (22)	14 (11)	21 (23)	26 (50)	$\chi^2(1) = 32.826$, $p < .001$[e]	
On a psychotropic, *n* (%)	92 (34)	62 (48)	29[f] (32)	1 (2)		
Race/ethnicity, *n* (%)						
White	132 (49)	75 (58)	22 (24)	35 (70)	$\chi^2(1) = 34.849$, $p < .001$[g]	
Hispanic/Latin(x)	57 (21)	18 (14)	35 (38)	4 (8)		
African-American	37 (14)	11 (9)	21 (23)	5 (10)		
Other	48 (18)	25 (19)	15 (16)	8 (15)		
Highest family education, *n* (%)						
Graduate/professional	132 (50)	70 (55)	21 (25)	41 (80)		
College	59 (23)	32 (25)	19 (23)	8 (16)		
Some college	29 (11)	15 (12)	13 (16)	1 (2)		
High school	28 (11)	8 (6)	19 (23)	1 (2)		
<High school	17 (6)	4 (3)	12 (14)	1 (2)		
EFCT raw score						
Total[abc]	7.8 (3.6)	9.2 (3.1)	8.0 (3.0)	3.8 (2.7)	$F(2, 14614.69) = 58.90$, $p < .001$	.31
(min–max)	(0–16)	(3–16)	(2–14)	(0–15)		
Flexibility[ac]	3.3 (2.3)	4.0 (2.1)	3.5 (2.0)	1.3 (1.8)	$F(2, 6214.99) = 31.0$, $p < .001$	.20
(min–max)	(0–8)	(0–8)	(0–8)	(0–7)		
Planning[abc]	4.4 (2.0)	5.2 (1.9)	4.5 (1.7)	2.4 (1.5)	$F(2, 6709.51) = 43.43$, $p < .001$	.26
(min–max)	(0–8)	(0–8)	(0–8)	(0–8)		
WASI						
FSIQ[abc] (standard score)	103.9 (16.0)	105.2 (16.2)	96.7 (14.1)	113.2 (13.1)	$F(2, 1.03e+30) = 21.31$, $p < .001$	.15
(min–max)	(72–151)	(75–151)	(72–134)	(86–140)		
Matrix reasoning[bc](T-score)	54.1 (10.1)	55.4 (9.8)	50.2 (9.9)	57.8 (8.8)	$F(2, 589.63) = 9.41$, $p < .001$	.09
(min–max)	(24–78)	(32–76)	(24–78)	(33–70)		
Vocabulary[ac](T-score)	52.5 (11.5)	51.0 (11.1)	49.5 (10.6)	61.4 (9.3)	$F(2, 5962.33) = 23.43$, $p < .001$	.15
(min–max)	(20–80)	(20–78)	(32–80)	(44–80)		
D-KEFS scale scores						
Category fluency	11.2 (3.8)	11.1 (4.2)	11.1 (3.2)	11.7 (3.7)	$F(2, 963.09) = .43$, $p = .65$	.01
(min–max)	(3–19)	(3–19)	(4–18)	(5–19)		
Switch accuracy[ac]	9.0 (3.2)	8.7 (3.0)	8.3 (3.2)	11.1 (2.8)	$F(2, 289.81) = 10.38$, $p < .001$	.11
(min–max)	(1–19)	(1–15)	(1–15)	(4–19)		
BRIEF T-scores						
Shift[abc]	62.5 (15.2)	69.2 (13.0)	63.3 (12.8)	44.3 (7.9)	$F(2, 65418.90) = 76.72$, $p < .001$	.37
(min–max)	(36–96)	(36–96)	(36–89)	(36–76)		
Organization/plan[ac]	60.4 (14.1)	65.1 (12.3)	62.3 (12.8)	45.2 (9.1)	$F(2,115193.10) = 52.28$, $p < .001$	.28
(min–max)	(33–86)	(35–84)	(35–86)	(33–76)		
GEC[abc]	62.4 (14.1)	68.1 (11.3)	64.3 (11.9)	44.6 (8.9)	$F(2, 2, 1380.89) = 71.43$, $p < .001$	.39
(min–max)	(33–88)	(36–88)	(33–86)	(33–78)		
ADHD rating scale T-scores						
Inattention[ac]	63.6 (14.6)	67.4 (12.7)	67.8 (13.4)	46.9 (7.9)	$F(2, 643.02) = 45.82$, $p < .001$	.39
(min–max)	(38–102)	(41–100)	(38–102)	(38–72)		
Hyperactivity[ac]	61.8 (15.9)	64.3 (14.9)	66.8 (15.4)	46.3 (7.3)	$F(2, 603.18) = 30.04$, $p < .001$	.23
(min–max)	(39–119)	(39–119)	(39–113)	(39–76)		
SCQ raw score						
Total score[ab]	10.0 (7.9)	13.7 (8.0)	8.3 (6.0)	8.3 (5.5)	$F(2, 452.26) = 30.11$, $p < .001$	.24
(min–max)	(0–35)	(0–35)	(0–31)	(0–25)		

Notes: ADOS, Autism Diagnostic Observation Scale; ASD, Autism Spectrum Disorder; D-KEFS, Delis–Kaplan Executive Function System; EFCT, Executive Function Challenge Task; BRIEF, Behavior Rating Inventory of Executive Function; FSIQ, Full-Scale IQ; GEC, Global Executive Composite; RRB, Restricted/Repetitive Behavior, SCQ, Social Communication Questionnaire; TD, Typically Developing Youth; WASI, Wechsler Abbreviated Scale of Intelligence.

Post hoc comparisons reveal significant ($p < .05$) differences between: [a]ASD *versus* TD, [b]ASD *versus* ADHD, and [c]ADHD *versus* TD.

[d] The degrees of freedom reported here reflect the pooling step in multiple imputation.

[e] This comparison indicates that males are over-represented in ADHD and ASD relative to TD groups.

[f] Medication data are missing for three participants in the ADHD group.

[g] Due to uneven group sizes, race/ethnicity was collapsed to white *versus* nonwhite. Youth with ADHD were more likely to be racial/ethnic minorities than TDs or youth with ASD.

Given that a substantial portion of children diagnosed with ASD have a pre-existing ADHD diagnosis (Miodovnik, Harstad, Sideridis, & Huntington, 2015), the clinical diagnosticians characterizing participants were alert to participants presenting with potential signs of ASD and without prior clinical diagnosis. Such individuals were directly assessed using the ADOS and the Social Communication Questionnaire. Any children who met criteria for ASD based on expert clinical judgment utilizing these measures were designated as having ASD. In addition, a large proportion of autistic children also meet criteria for ADHD (41–78%; Murray, 2010). This sample is no exception. ASD was

https://doi.org/10.1017/S135561772000017X Published online by Cambridge University Press

accepted as the primary diagnosis, and there was not a formal evaluation of ADHD diagnoses in the participants with established ASD diagnoses. Based on the ADHD Rating Scale (DuPaul, Power, Anastopoulos, & Reid, 1998), there were high rates of parent-reported ADHD symptoms in children with ASD (40% of autistic participants were elevated on Inattention and 35% on Hyperactivity). In comparison, parent ratings on the ADHD Rating Scale in the ADHD group indicated that 51% of the participants received a T-score above 70 on the Inattention Scale and 45% on the Hyperactivity Scale.

Measures

The EFCT

The EFCT (Kenworthy et al., 2014) measures EF skills in the context of a semi-structured, interactive 20-min task that challenges children to be flexible and planful during four activities with an examiner (puzzle, modeling, drawing, and scenarios). See Supplemental Materials for the EFCT tasks and scoring criteria. Each task requires the child to verbally generate a plan and respond to two provocative flexibility challenges (e.g., participant is interrupted in the middle of constructing a clay figure and told to trade their figure with the examiner's figure). The EFCT has task-specific, precise behavioral markers to guide scoring on a three-point scale. The EFCT yields flexibility and planning raw scores (scores range from 0 to 8; higher scores indicate greater impairment) as well as a total raw score. There are two parallel forms of the EFCT (A and B). Form B was given to a small subset of youth with ASD and ADHD.

Inter-rater reliability was examined among trained research assistants, one school staff professional, psychology trainees, and clinical psychologists via percent agreement during a two-step process: examiners (1) received approximately 1 hr of didactic instruction in administration and scoring from one of the two primary authors (LA and LK); (2) watched several archival videos of EFCT administrations by reliable examiners; and (3) completed two in-person EFCTs with 80% reliability with one of the authors (LA, KD, LK, AV, or AR). Eighteen examiners were trained to administer the EFCT, and most achieved reliability within three administrations. One person did not achieve reliability following this process and thus did not administer the EFCT independently.

Validation measures

Participants and their parents were also administered measures of hypothesized convergent validity: Delis–Kaplan Executive Function System (D-KEFS; Delis, Kaplan, & Kramer, 2001), Category Fluency and Category Switching; Behavior Rating Inventory of Executive Function – Parent Report (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000), and ADOS Restricted/Repetitive Behavior raw score. Divergent validity was investigated using the ADOS Social Affect (SA) raw score in the participants with ASD and the Wechsler Abbreviated Scale of Intelligence first or second edition Vocabulary and Matrix Reasoning score in all participants.

Data Analytic Plan

Following the imputation of missing data using multiple imputation by chained equations ($m = 40$, Van Buuren & Groothuis-Oudshoorn, 2011), psychometrics (i.e., factor structure, reliability, and precision) were tested with confirmatory factor analysis (CFA), Cronbach's α, and reliable change index (RCI; Jacobson & Truaxx, 1991). Construct validity was examined with ANOVA, correlations, and hierarchical regression.

RESULTS

Hypothesis 1: The EFCT will Separate into Flexibility and Planning Factors

The eight-item EFCT was designed to measure two subdomains of EF – planning and flexibility. CFA was fit using diagonally weighted least squares on ordered categorical items in the lavaan (Rosseel, 2012). The two-factor model demonstrated excellent fit ($\chi^2(19) = 27.38$, $p = .10$, Comparative Fit Index (CFI) = .98, Root Mean Square Error of Approximation (RMSEA) = .04, Standardized Root Mean Squared Residual (SRMR) = .04). In contrast, a unidimensional model in which all items loaded on an EF factor displayed poorer fit ($\chi^2(20) = 55.06$, $p < .001$, CFI = .90, RMSEA = .08, SRMR = .05). The two-dimensional model fit significantly better, $\chi^2(1) = 23.96$, $p < .001$.

Measurement invariance was tested across diagnostic groups by a series of more restrictive models – configural invariance (i.e., no constraints, same factor structure), weak invariance (i.e., loadings constrained to be equal), and strong invariance (i.e., loadings and intercepts constrained to be equal). Weak invariance indicates that the EFCT construct (factors) has the same meaning across groups. Strong invariance indicates that the item scores are equivalent across groups. The configural model fit for all diagnostic groups, $ps > .30$. Across diagnostic groups, the EFCT demonstrated weak ($\chi^2(8) = 11.18$, $p = .19$, ΔCFI < .001, ΔRMSEA < .001, ΔSRMR = .01) but not strong invariance ($\chi^2(12) = 41.33$, $p < .001$, ΔCFI = .17, ΔRMSEA = .07, ΔSRMR = .02), indicating that EFCT factors were consistent across groups, but average EFCT scores varied by group.

Hypothesis 2: Forms A/B will be Equivalent

Exploratory tests of measurement invariance of Forms A/B of the EFCT were conducted. First, measurement invariance was examined across Form A ($n = 225$ with ASD, ADHD, or TD) and Form B ($n = 49$ with ASD or ADHD). The same series of more restrictive models were fit examining configural, weak, and strong invariance. The configural model fit both Form A and Form B, $ps > .11$. Forms A/B demonstrated

https://doi.org/10.1017/S135561772000017X Published online by Cambridge University Press

Table 2. Correlation matrix of executive control measures in all participants ($n = 274$) with *a priori* hypotheses indicated by arrows

		EFCT								
		Total			Planning			Flexibility		
		H_x[a]	r[b]	95% CI[c]	H_x	r	95% CI	H_x	r	95% CI
EFCT	Planning	↑	.82*	.78, .86				↑		
	Flexibility	↑	.86*	.83, .89	↑	.43*	.32, .52			
D-KEFS	Category fluency	↓	−.14	.02, −.29	↓	−.18	−.02, −.33		−.07	.09, −.22
	Switch accuracy	↓	−.29*	−.13, −.43		−.28*	−.12, −.43	↓	−.21	−.06, −.35
BRIEF	Shift	↑	.40*	.30, .50		.33*	.21, .43	↑	.36*	.24, .46
	Plan/Org.	↑	.33*	.21, .44	↑	.29*	.18, .40		.26*	.14, .38
	GEC	↑	.42*	.31, .52	↑	.36*	.24, .46	↑	.35*	.23, .46
ADOS	Social Affect	ø	.18	−.04, .37	ø	.02	−.18, .21	ø	−.01	.19, −.22
($n = 129$, ASD)	RRB	↑	−.01	.22, −.24		−.14	.11, −.36	↑	.10	−.10, .30
WASI	Vocabulary	ø	−.44*	−.33, −.54	ø	−.47*	−.36, −.56	ø	−.29*	−.17, −.41
	Matrix reasoning	ø	−.21	−.07, −.34	ø	−.22	−.09, −.35	ø	−.14	.00, −.27
Demographics	Age		−.29*	−.17, −.39		−.28*	−.17, −.39		−.20	−.08, −.32
	Highest parental education		.12	−.03, .26		.11	−.04, .25		.09	−.07, .25

Notes: ADOS, Autism Diagnostic Observation Schedule; BRIEF, Behavior Rating Inventory of Executive Function; GEC, Global Executive Composite; D-KEFS, Delis–Kaplan Executive Function System; EFCT, Executive Function Challenge Task; RRB, Restricted/Repetitive Behavior; WASI, Wechsler Abbreviated Scale of Intelligence.

[a] ↑↓ Arrows represent *a priori* hypothesized (H_x) directional relationships (convergent validity). ∅ represents *a priori* hypothesis (H_x) of no significant correlation (divergent validity).

[b] r = Pearson's r, except for the ADOS variables, for which statistic is Spearman's Rho.

[c] CI = Confidence interval; (95% CI) are unadjusted.

*Significant correlation, after *Holms-Stepdown* corrected $p < .05$, for all possible correlations in the complete correlation matrix.

weak ($\chi^2(4) = 5.25$, $p = .26$, ΔCFI = .006, ΔRMSEA = .001, ΔSRMR = .004) and strong invariance ($\chi^2(6) = 5.39$, $p = .49$, ΔCFI = .002, ΔRMSEA = .004, ΔSRMR = .002), indicating that the factor structure was consistent across forms, and participants performed similarly across forms. In a separate set of data, Form A ($n = 96$) and Form B ($n = 75$) were administered to youth with ADHD and ASD. Fit indices indicated that the configural model fit Form A and Form B reasonably well. The two EFCT forms demonstrated strong invariance ($\chi^2(6) = 8.34$, $p = .21$, ΔCFI = .02, ΔRMSEA = .002, ΔSRMR = .002). In summary, preliminary evidence suggests that the two forms of the EFCT may be interchangeable.

Hypothesis 3: The EFCT will Demonstrate Internal Reliability

Cronbach's α estimated the internal consistency in the total sample. Internal consistency was in the adequate (planning = .78) to good range (flexibility = .81, total = .84) given the brevity of the scale. The RCI indicates how much a score must change for the change to be more than measurement error. The 90% RCI was 2.7 for total, 2.0 for flexibility, and 1.9 for planning. These RCIs suggest that relatively small changes in scores indicate reliable idiographic change.

Hypothesis 3a: The EFCT will Discriminate Youth with ASD or ADHD from TD Youth

A series of ANOVAs compared EFCT total, planning, and flexibility scale scores among ASD, ADHD, and TD youth. ANOVAs comparing the groups on key demographic variables revealed that they differed significantly in age, FSIQ, gender, race/ethnicity, and parent education level. See Table 1. However, ANCOVA sensitivity analyses controlled for youth's sex, age, parent education level, and verbal ability and indicated no change in the pattern of significant findings (see Supplementary Material B). On the EFCT, ASD youth had higher (worse) total scores than ADHD youth (Cohen's d = .41) and TD youth (Cohen's d = 1.81) and ADHD youth had higher total scores than TD youth (Cohen's d = 1.45). ASD and ADHD youth had higher flexibility scores than TD youth (Cohen's d = 1.34 and 1.35, respectively). ASD youth had higher planning scores than ADHD youth (Cohen's d = .38) and TD youth (Cohen's d = 1.56) and ADHD youth had higher planning scores than TD youth (Cohen's d = 1.29).

Hypothesis 3b and 3c: The EFCT will Demonstrate Convergent and Discriminant Validity

Table 2 displays bivariate correlations. The Holm–Bonferroni procedure controlled the error rate (Holm, 1979). The EFCT scores were weakly associated with age and moderately with verbal ability such that older youth or youth with better verbal ability tended to have better EF. EFCT scores were not associated with parental education or the quality of a youth's social interaction. Sensitivity analyses examined the unique relationship between the EFCT and hypothesized variables controlling for age and verbal ability with hierarchical

https://doi.org/10.1017/S135561772000017X Published online by Cambridge University Press

regressions (Supplemental Material Table 2). Consistent with hypotheses, the EFCT scales, but not the D-KEFS scales, were moderately associated with the BRIEF, and this remained after controlling for covariates. Contrary to our hypothesis, the D-KEFS was not associated with the EFCT after controlling for covariates.

DISCUSSION

The EFCT shows initial promise as a reliable, valid measure of EF with the potential to address gaps in current pediatric EF measurement, including the need for treatment-sensitive measures that can be administered in a standardized and masked fashion and incorporate "hot" EF demands. It is related to parent-reported everyday EF problems, indicating its potential as an ecologically valid, performance-based measure.

The EFCT demonstrates the hypothesized two-factor structure of flexibility and planning dimensions, and there is preliminary evidence of measurement invariance, as well as adequate to good internal consistency and strong discriminant validity between youth with ASD or ADHD and TD youth. EFCT total and planning scores are sensitive to the development of EF abilities, as demonstrated by their negative correlation with age, but are not related to socioeconomic status (i.e., parental education). The EFCT has previously demonstrated sensitivity to treatment (Kenworthy et al., 2014), and the RCI scores calculated in this study indicate that it can capture idiographic change with precision. The parallel forms of the EFCT show preliminary evidence of invariance, further indicating its promise for use in treatment trials. Performance on the EFCT discriminates youth with ASD from those with ADHD or TD, as well as youth with ADHD from those with TD, even after controlling for group differences in age, gender, FSIQ, and parent education.

The construct validity of the EFCT is supported by its positive associations with parent-reported EF, which were not found in the other performance-based EF task used in this study. Consistent with previous reviews demonstrating no significant relationship between traditional performance-based EF tasks and informant report measures of EF (Toplak et al., 2013), parent-reported EF problems were not significantly related to D-KEFS Fluency scores. In ASD and ADHD, rating scales of EF are valued for their capacity to predict key outcomes, such as impairment in major life activities (Barkley & Fischer, 2011) and adaptive skills in youth (Pugliese et al., 2016). In contrast, traditional performance-based tasks have been identified as, by their nature, ill-suited to replicate real-world EF demands, because they make expectations explicit (e.g., "work as quickly as you can") instead of implicit. As such, they provide a good understanding of a child's *potential* EF when structured supports are provided but often fail to capture the child's *actual* EF skills in everyday settings because youth are expected to make choices, decisions, and plans without overt instructions (Toplak et al., 2013). Even tasks with face validity, which are designed to look like real-world EF challenges, are not guaranteed to predict real-world outcomes, unless the demands of the EF task actually resemble real-world demands (Kraybill, Thorgusen, & Suchy, 2013). The EFCT was developed to replicate the EF demands youth encounter every day to flexibly plan, respond to setbacks, and achieve goals while interacting with others.

The hypothesized relationship of the EFCT to a traditional performance-based EF task (D-KEFS Fluency) was not supported in this study. Furthermore, despite being related at the global level, correlation and regression analyses did not indicate differential relationships between purported measures of flexibility or planning within the BRIEF and the EFCT. This could reflect a lack of specificity in the EFCT, other measures, or in the construct of EF itself in youth (Lee, Bull, & Ho, 2013). Because the EFCT flexibility and planning scores are differentially responsive to treatment, and the two-factor solution was a better fit than the unitary one in the CFA, the question of whether the EFCT captures fractionated components of EF in addition to the unitary construct should be further investigated in the EFCT and a broader range of EF performance tasks.

Regarding divergent validity, the EFCT was not significantly correlated with ADOS SA score, as measured in participants with ASD only. This is a key initial indication that the EFCT scoring system is not inadvertently assessing social capacities and difficulties understanding the examiner's expectations (White, 2013). Nor was it related to the Wechsler Matrix Reasoning score, distinguishing it from a measure of visual pattern recognition. EFCT scores were correlated with vocabulary scores, reflecting the confounding of language and executive demands, which limits the EFCT's utility with youth with impaired language. This problem is, unfortunately, not unique to the EFCT, as EF tests often are criticized for task impurity, whereby nonexecutive abilities drive performance on tests purported to measure EF (Miyake & Friedman, 2012).

A limitation of this investigation is that it reports on data that are aggregated across multiple samples, collected for different purposes (treatment and phenotyping trials). While this sample of convenience indicates that the EFCT has adequate psychometric properties, even across diverse youth and recruitment strategies, further planned investigations are required with: a wider range of validation measures, more comprehensive assessment of co-morbid conditions in clinical samples, more TD youth, who are more closely matched to the clinical samples (e.g., IQ, SES), and additional psychometric investigations, especially test–retest stability. A limitation of the measure as created is that the strong verbal mediation of the measure means it would need to be adapted for use in an intellectually disabled population. Nonetheless, the EFCT has promise as a tool that taps real-world EF ability, which is fundamental to mental health and other outcomes, and thus a key target for intervention and biomarker research.

https://doi.org/10.1017/S135561772000017X Published online by Cambridge University Press

ACKNOWLEDGEMENTS

We thank the children and families who participated in this study.

This work was supported by the Gudelsky Family Foundation (L.K.), the Patient-Centered Outcomes Research Institute (L.K. and L.A., AD-1304-7379), and the National Institutes of Health (L.K. and L.A., NIMH R34MH083053-01A2), (V.G., NIH1U54HD090257), (C.P., T32 HD046388-01A2 and K23MH110612). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

CONFLICTS OF INTEREST

L.K. receives royalties sales of the Behavior Rating Inventory of Executive Function. There are no other conflicts of interest, financial, or otherwise for the authors involved directly or indirectly with this manuscript.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S135561772000017X

References

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.

Barkley, R.A., & Fischer, M. (2011). Predicting impairment in major life activities and occupational functioning in hyperactive children as adults: Self-reported executive function (EF) deficits versus EF tests. *Developmental Neuropsychology*, *36*, 137–161.

Burgess, P.W., Alderman, N., Forbes, C., Costello, A., Coates, L.M.A., Dawson, D.R., Anderson, N.D., Gilbert, S.J., Dumontheil, I., & Channon, S. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society*, *12*, 194–209.

Chevignard, M., Pillon, B., Pradat-Diehl, T., Taillerfer, C., Rousseau, S., Le Bras, C., & Dubois, B. (2000). An ecological approach to planning dysfunction: Script execution. *Cortex*, *36*, 649–669.

Delis, D.C., Kaplan, E., & Kramer, J.H. (2001). *The Delis-Kaplan Executive Function System: Examiner's Manual*. San Antonio: Psychological Corporation.

DuPaul, G.J., Power, T.J., Anastopoulos, A.D., & Reid, R. (1998). *ADHD Rating Scale IV: Checklists, Norms, and Clinical Interpretation*. New York: Guilford Press.

Friedman, N.P. & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186–204.

Gioia, G.A., Isquith, P.K., Guy, S., & Kenworthy, L. (2000). *BRIEF: Behavior Rating Inventory of Executive Function*. Odessa, FL: Psychological Assessment Resources.

Gioia, G.A., Isquith, P.K., Kenworthy, L., & Barton, C. (2002). Executive signatures: Profiles of everyday executive function in acquired and developmental disorders. *Child Neuropsychology*, *8*, 121–137.

Goldsmith, H.H., & Rothbart, M.K. (1996). *Prelocomotor and Locomotor Laboratory Temperament Assessment Battery, Lab-TAB; version 3.0*. Technical Manual, Department of Psychology, University of Wisconsin, Madison, WI.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Holmes-Bernstein, J., & Waber, D.P. (1990). Developmental neuropsychological assessment. In Boulton, A. A., Baker, G. B., Hiscock, M. (Eds.), *Neuropsychology: Neuromethods* (Vol. *17)* (pp. 311–371). Totowa, NJ: Humana Press.

Jacobson, N.S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Kenworthy, L., Anthony, L.G., Naiman, D.Q., Cannon, L., Wills, M.C., Luong-Tran, C., Adler Werner, M., Alexander, K.C., Strang, J., Bal, E., Sokoloff, J., & Wallace, G.L. (2014). Randomized controlled effectiveness trial of executive function intervention for children on the autism spectrum. *Journal of Child Psychology and Psychiatry*, *55*, 374–383.

Kraybill, M.L., Thorgusen, S.R., & Suchy, Y. (2013). The push-turn-taptap task outperforms measures of executive functioning in predicting declines in functionality: Evidence-based approach to test validation. *The Clinical Neuropsychologist*, *27*, 238–255.

Lee, K., Bull, R., & Ho, R.M.H. (2013). Developmental changes in executive functioning. *Child Development*, *84*, 1933–1953.

Miodovnik, A., Harstad, E., Sideridis, G., & Huntington, N. (2015). Timing of the diagnosis of attention-deficit/hyperactivity disorder and autism spectrum disorder. *Pediatrics*, *136*, 830–837.

Miyake, A. & Friedman, N.P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*, 8–14.

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.

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H., Houts, R., Poulton, R., Roberts, B.W., Ross, S., Sears, M.R., Thomson, W.M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academic of Sciences*, *108*, 2693–2698.

Murray, M.J. (2010). Attention-deficit/hyperactivity disorder in the context of autism spectrum disorders. *Current Psychiatry Reports*, *12*, 382–388.

Ozga, J.E., Povroznik, J.M., Engler-Chiurazzi, E.B., & Vonder Haar, C. (2018). Executive (dys)function after traumatic brain injury: Special considerations for behavioral pharmacology. *Behavioural Pharmacology*, *29*, 617–637.

Poland, S.E., Monks, C.P., & Tsermentseli, S. (2014). Cool and hot executive function as predictors of aggression in early childhood: Differentiating between the function and form of aggression. *British Journal of Developmental Psychology*, *34*, 181–197.

Prencipe, A., Kesek, A., Cohen, J., Lamm, C., Lewis, M.D., & Zelazo, P.D. (2011). Development of hot and cool executive

https://doi.org/10.1017/S135561772000017X Published online by Cambridge University Press

function during the transition to adolescence. *Journal of Experimental Child Psychology*, *108*, 621–637.

Pugliese, C.E., Anthony, L.G., Strang, J.F., Dudley, K., Wallace, G.L., Naiman, D.Q., & Kenworthy, L. (2016). Longitudinal examination of adaptive behavior in autism spectrum disorders: Influence of executive function. *Journal of Autism and Developmental Disorders*, *46*, 467–477.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Snyder, H.R., Miyake, A., & Hankin, B. (2015). Advancing understanding of executive function impairments and psychopathology: Bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology*, *6*, 1–24.

Strang, J.F., Anthony, L.G., Yerys, B., Hardy, K.K., Wallace, G.L., Armour, A.C., Dudley, K., & Kenworthy, L. (2017). The flexibility scale: Development and preliminary validation of a cognitive flexibility measure in children with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *47*, 2502–2518.

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67.

White, S.J. (2013). The triple I hypothesis: Taking another('s) perspective on executive dysfunction in autism. *Journal of Autism and Developmental Disorders*, *43*, 114–121.

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

Zelazo, P.D. & Cunningham, W.A. (2007). Executive function: Mechanisms underlying emotion regulation. In J.J. Gross (Ed.), *Handbook of Emotion Regulation* (pp. 135–158). New York, NY: Guilford.

https://doi.org/10.1017/S135561772000017X Published online by Cambridge University Press